

## P

## Partial Differential Equations that Lead to Solitons

DOĞAN KAYA

Department of Mathematics, Firat University,  
Elazığ, Turkey

### Article Outline

[Definition of the Subject](#)

[Introduction](#)

[Some Nonlinear Models that Lead to Solitons](#)

[Future Directions](#)

[Bibliography](#)

### Definition of the Subject

In this part, we introduce the reader to a certain class of nonlinear partial differential equations which are characterized by solitary wave solutions of the classical nonlinear equations that lead to solitons. The classical nonlinear equations of interest show the existence of special types of traveling wave solutions which are either solitary waves or solitons. In this study, we will review a few solutions arising from the analytic work of the Korteweg–de Vries (KdV) equations, the generalized regularized long-wave RLW equation, Kadomtsev–Petviashvili (KP) equation, the Klein–Gordon (KG) equation, the Sine-Gordon (SG) equation, the Boussinesq equation, Pochhammer–Chree (PC) equation and the nonlinear Schrödinger (NLS) equation, the Fisher equation, Burgers equation, the Korteweg–de Vries Burgers’ equation (KdVB), the two-dimensional Korteweg–de Vries Burgers’ (tdKdVB), the potential Kadomtsev–Petviashvili equation, the Kawahara equation, Generalized Zakharov–Kuznetsov (gZK) equation, the Sharma–Tasso–Olver equation, and the Cahn–Hilliard equation.

### Introduction

Nonlinear phenomena play a crucial role in applied mathematics and physics. Calculating exact and numerical solutions, in particular, traveling wave solutions, of nonlinear PDEs in mathematical physics plays an important role in soliton theory. Moreover, these equations are mathematical models of complex physical occurrences that arise in engineering, chemistry, biology, mechanics, and physics.

In this work, we give a brief history of the above-mentioned nonlinear equations and how this type of equation has led to the soliton solutions; we then present an introduction to the theory of solitons. Soliton theory is an important branch of applied mathematics and mathematical physics. In the last decade this topic has become an active and productive area of research, and applications of the soliton equations in physical cases have been considered. These have important applications in fluid mechanics, nonlinear optics, ion plasma, classical and quantum fields’ theories etc.

The best introduction to the soliton is that contained in J. Scott Russell’s (a Scottish naval engineer) seminal 1844 report to the Royal Society. Scott Russell’s report titled “Report on Waves” [62] was presented to the Royal Society in the 18th Century and he wrote:

“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 grad-

ually diminished, and after a chase of one or two miles I lost it in the windings of the channel . . .”

Scott Russell’s experimental observations in 1834 were followed by the theoretical scientific work of Lord Rayleigh and Joseph Boussinesq around 1870 [60] and finally, two Dutchmen, Korteweg and de Vries, developed a nonlinear partial differential equation to model the propagation of shallow water waves applicable to the situation in 1895 [47]. This work was really what Scott Russell fortuitously witnessed. This famous classical equation is known simply as the KdV equation. Korteweg and de Vries published a theory of shallow water waves which reduced Russell’s observations to its essential features. The nonlinear classical dispersive equation was formulated by Korteweg and de Vries in the form

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} + \varepsilon \frac{\partial^3 u}{\partial x^3} + \gamma u \frac{\partial u}{\partial x} = 0$$

where  $c$ ,  $\varepsilon$ ,  $\gamma$  are physical parameters. This equation plays a key role in soliton theory.

In the late 1960s, Zabusky and Kruskal numerically studied and then analytically solved the KdV equation [79]. They came to the result that stable pulse-like waves could exist in a problem described by the KdV equation from their numerical study. A remarkable aspect is that the discovered solitary waves retain their shapes and speeds after collision. Zabusky and Kruskal called these waves solitons as they resemble particles in nature. In 1967, this numerical development was placed on a settled mathematical basis with Gardner et al.’s discovery of the inverse-scattering-transform method [20].

The soliton concept is related with solutions for nonlinear partial differential equations. The soliton solution of a nonlinear equation usually is used a single wave. If there are several soliton solutions, these solutions are called solitons. On the other hand, if a soliton is separated infinitely from another soliton, this soliton is called a single wave. Besides, a single wave solution can’t be  $\sec h^2$  function for equations other than nonlinear equations, such as the KdV equation. But this solution can be  $\text{sech}$  or  $\tan^{-1}(e^{\alpha x})$ .

At this stage, we could ask what the definition of the soliton solutions is? It is not easy to define the soliton concept. Wazwaz [69] describes solitons as solutions of nonlinear differential equations as follows:

- (i) A long and shallow water wave should not lose its permanent form;
- (ii) A long and shallow water wave of the solution is localized, meaning either the solutions decay exponentially to zero such as in the solitons admitted by the

KdV equation, or approach a constant at infinity such as solitons provide by the SG equation;

- (iii) A long and shallow water wave of the solution can interact with other solitons preserving its character.

There is also a more formal definition of this concept, but these definitions require substantial mathematics [17]. On the other hand, the phenomena of the solitons doesn’t always follow these three properties. For example, there is the concept of “light bullets” in the subject of nonlinear optics which are often called solitons despite losing energy during interaction. This idea can be found at an internet web page of the [Simon Fraser University](#) British Columbia, Canada [51].

### Some Nonlinear Models that Lead to Solitons

In this section, we will deal with the fundamental ideas and some nonlinear partial differential equations that lead to solitons. These equations are the result of a huge amount of research work and physical implementations which focus on the most diverse and active areas of applied mathematics and mathematical physics.

*Example 1* In various fields of science and engineering, nonlinear evaluation equations, as well as their analytic and numerical solutions, are of fundamental importance. One of the most attractive and surprising wave phenomena is the creation of solitary waves or solitons. It was approximately two centuries ago, that an adequate theory for solitary waves was developed, in the form of a modified wave equation known as the KdV equation [14,15,16,47,69,75]. The third-order KdV equation is a classical nonlinear partial differential equation originally formulated to model shallow water flow [47].

Herein, we are particularly interested in the original third-order KdV equation [47] given by

$$u_t + u u_x + \alpha u_{xxx} = 0,$$

respectively, and the generalized form

$$u_t + u^p u_x + \alpha u_{xxx} = 0,$$

where  $p$  models the dispersion and subscripts in  $t$  and  $x$  denote partial derivatives with respect to these independent variables. The study of long waves, tides, solitary waves, and related phenomena, leads to an equation referred to as the generalized KdV equation. If  $p = 0$ ,  $p = 1$ , and  $p = 2$ , this equation becomes the linearized KdV, nonlinear KdV, and modified KdV equations, respectively [14,15,16,69,75]. The modified KdV equation has

also found applications in many areas of physical situations, for instance the equation describes nonlinear acoustic waves in an anharmonic lattice [78].

**Example 2** Consider the solution  $u(x, t)$  of the generalized RLW equation

$$u_t + u_x + \alpha u_x^v - \beta u_{xxt} = 0,$$

where  $v$  is a positive integer, and  $\alpha$  and  $\beta$  are positive constants. The generalized RLW equation was first put forward as a model for small-amplitude long waves on the surface of water in a channel by Peregrine [37,57] and later by Benjamin et al. [2]. In physical situations such as unidirectional waves propagating in a water channel, long-crested waves in near-shore zones, and many others, the generalized RLW equation serves as an alternative model to the KdV equation [4,5].

**Example 3** The nonlinear KG equations in one-dimensional space

$$u_{tt} - u_{xx} + \frac{dV}{du} = 0,$$

where  $V = V(u)$  is a general nonlinear function of  $u$ , but not its derivatives. This equation is the most natural nonlinear generalization of the wave equation and first arose in a mathematical context, with  $V(u) = \exp(u)$ , in the theory of surfaces of constant curvature. In addition, this equation appears in many different fields of application. For instance, a polynomial nonlinearity can be used as a model field theory, while a  $\cos u$  term yields the sine-Gordon equation and so on [14,15,16,69,75]. We will consider a particular case of equation KG, the so-called sine-Gordon nonlinear hyperbolic equation, which has the form

$$u_{tt} - cu_{xx} + \kappa \sin(u) = 0; \quad -\infty < x < \infty, \quad t > 0,$$

where  $c$  and  $\kappa$  are constants. The sine-Gordon equation is firstly used in the work of differential geometry and for investigation of the propagation of a 'slip' dislocation in crystals [15].

The generalized one-dimensional KG equation

$$u_{tt} - ku_{xx} + b_1 u + b_2 u^{m+1} + b_3 u^{2m+1} = 0,$$

is given in [80]. The KG equation represents a nonlinear model of longitudinal wave propagation of elastic rods when  $m = 1$  [12].

**Example 4** In this example, we consider the generalized Boussinesq-type equation [7]

$$u_{tt} - u_{xx} + \delta u_{xxxx} = -(\psi(u))_{xx}; \\ -\infty < x < \infty, \quad t > 0,$$

where  $\delta \geq 0$  is constant,  $\psi(u) = |u|^{\alpha-1} u$  and  $\alpha > 1$ . This equation represents a generalization of the classical Boussinesq equation which arises in the modeling of nonlinear strings. The Boussinesq equation describes in the continuous limit the propagation of waves in a one-dimensional nonlinear lattice and the propagation of waves in shallow water [6,7,15,77]. A proof of the local well-posedness of the Boussinesq equation has been showed by Bona and Sachs [6]. In [13] the authors noticed that the Boussinesq equation admits solitary solutions and the existence of solitary wave solutions illustrates the perfect balance between dispersion and the nonlinearity of the Boussinesq equation. For the  $\alpha$  integer, solitary-wave solutions have been shown to be stable under some restrictions on the wave speed by Bona and Sachs [6].

Scott Russell's study [62] of solitary water waves motivated the development of nonlinear partial differential equations for the modeling of wave phenomena in fluids, plasmas, elastic bodies, etc. The Boussinesq equation is an important model that approximately describes the propagation of long waves on shallow water like the other Boussinesq equations (with  $u_{xxtt}$ , instead of  $u_{xxxx}$ ). This equation was first deduced by Boussinesq [7]. In the case  $\delta > 0$  this equation is linearly stable and governs small nonlinear transverse oscillations of an elastic beam [15]. It is called the "good" Boussinesq equation, while the equation with  $\delta < 0$  received the name "bad" Boussinesq equation since it possesses linear instability.

**Example 5** Consider the PC equation

$$u_{tt} - \frac{1}{p} u_{xx}^\ell - u_{xx} - u_{ttxx} = 0,$$

where  $\ell$  indicates different material with different values of it. The PC equation is applied as a nonlinear model of longitudinal wave propagation of elastic rods [15,16,75]. In the work of Bogolubsky [75], the author obtained exact solitary wave solutions to the PC equation  $\ell = 2, 3, 5$ , respectively [15].

**Example 6** In this paper, we consider the Burgers equation

$$u_t + \varepsilon u u_x - \nu u_{xx} = 0$$

where  $\varepsilon$  and  $\nu$  are parameters and the subscripts  $t$  and  $x$  denote differentiation. The Burgers' equation is a model of flow through a shock wave in a viscous fluid [13] and in the Burgers' model of turbulence [9].

**Example 7** The Fisher equation

$$u_t - \nu u_{xx} = ku \left(1 - \frac{u}{\kappa}\right),$$

is well known in population dynamics, where  $\nu > 0$  is the diffusion constant,  $k > 0$  is the linear growth rate, and  $\kappa > 0$  is the carrying capacity of the environment. The right-hand side function  $ku(1 - \frac{u}{\kappa})$  represents a nonlinear growth rate [15]. This well-known equation was first proposed by Fisher [18] for a model of the advancement of a mutant gene in an infinite one-dimensional habitat. In recent years, the equation has been used as a basis for a wide variety of models for the spatial spread of genes in a population and for chemical wave propagation [18].

**Example 8** Let us consider the  $(1 + 1)$ -dimensional NLS equation with two higher-order nonlinear terms in the form

$$iu_t + \frac{1}{2}u_{xx} + \alpha |u|^\sigma u + \beta |u|^{2\sigma} u = 0,$$

where  $\sigma$  is a positive constant,  $\alpha$  and  $\beta$  are constant parameters [15,16,69].  $u = u(x, t)$  is a complex function that represents the complex amplitude of the wave form, the variable  $t$  should be interpreted as the normalized propagation distance,  $x$  the normalized transverse coordinate that represents a retarded time. The NLS equation, especially that with lower  $\sigma$  values, appears in various branches of contemporary physics [18,27,61,67,80,81] and has been extensively investigated for the cases of  $\sigma = 1$  and  $\sigma = 2$ , its various solutions have been obtained. The NLS equation also arises in some form of  $\alpha = 0$  and  $\sigma = 1$  in some other physical systems such as nonlinear optic [3,26,38], hydro magnetic and plasma waves [30,63] and the propagation of solitary waves in piezoelectric semiconductors [55].

**Example 9** In applied mathematics, the KP equation [36] is a nonlinear partial differential equation. It is also sometimes called the Kadomtsev–Petviashvili–Boussinesq equation. The KP equation is usually written as:

$$(u_t - 6uu_{xx} + u_{xxx})_x + 3\lambda^2 u_{yy} = 0,$$

where  $\lambda = \pm 1$ . This equation is implemented to describe slowly varying nonlinear waves in a dispersive medium [58]. The above written form shows that the KP equation is a generalization form of the KdV equation which is like the KdV equation; the KP equation is completely integrable. The KP equation was first discovered in 1970 by Kadomtsev and Petviashvili when they relaxed the restriction that the waves be strictly one-dimensional.

**Example 10** Consider the solution  $u(x, t)$  of the nonlinear partial differential equation

$$u_t + \varepsilon uu_x - \nu u_{xx} + \mu u_{xxx} = 0$$

where  $\varepsilon$ ,  $\nu$  and  $\mu$  are positive parameters. This equation is called the Korteweg–de Vries Burgers' equation (KdVB) which is derived by Su and Gardner [65]. KdVB is a model equation for a wide class of nonlinear systems in the weak nonlinearity and long wavelength approximations because it contains both damping and dispersion. KdVB has been constructed when including electron inertia effects in the description of weak nonlinear plasma waves. The KdVB equation possesses a steady-state solution which has been demonstrated to model weak plasma shocks propagating perpendicular to a magnetic field [24]. There are some other implementation places to use the KdVB equation. Examples are its use in the study of wave propagation through a liquid-filled elastic tube [34] and for a description of shallow water waves on a viscous fluid [21,35].

**Example 11** The nonlinear partial differential equation

$$(u_t + uu_x - qu_{xx} + pu_{xxx})_x + ru_{yy} = 0,$$

where  $p$ ,  $q$ ,  $r$  are real parameters. This equation is called the two-dimensional Korteweg–de Vries Burgers' (tdKdVB) equation. The tdKdVB equation is a model equation for a wide class of nonlinear wave models of fluids in an elastic tube, liquids with small bubbles and turbulence [52,53].

**Example 12** In this example we consider in the  $(2 + 1)$ -dimension the potential Kadomtsev–Petviashvili equation,

$$u_{xt} + \frac{3}{2}u_x u_{xx} + \frac{1}{4}u_{xxxx} + \frac{3}{4}u_{yy} = 0,$$

where the initial conditions  $u(x, 0, t)$  and  $u_y(x, 0, t)$  are given. Nonlinear phenomena play a crucial role in applied mathematics and physics. Obtaining the exact or approximate solutions of PDEs in physics and mathematics is important and it is still a hot spot to seek new methods to obtain new exact or approximate solutions [14,15,16]. Different methods have been put forward to seek various exact solutions of multifarious physical models described by nonlinear PDEs.

**Example 13** We consider the numerical solution to a problem involving a nonlinear partial differential equation of the form

$$u_t + u u_x + u_{xxx} - u_{xxxx} = 0,$$

this is called the Kawahara equation. The Kawahara equation occurs in the theory of magneto-acoustic waves in plasmas [39] and in the theory of shallow water waves with surface tension [29].



**Example 14** A Generalized Zakharov–Kuznetsov (gZK) equation [14,15,16,48,75] is proposed to understand the physical and scientific mechanisms in different physical and engineering problems. The gZK equation is as follows

$$u_t + b_1 u^\rho u_x + b_2 u^{2\rho} u_x + b_3 u_{xy} + b_4 u_{xxx} + b_5 u_{xyy} = 0,$$

where  $b_1, b_2, b_3, b_4 \neq 0, b_5 \neq 0$  and  $\rho \neq 0$  are constant parameters [48]. When  $b_1 = 6, b_2 = b_3 = 0, b_4 = b_5 = 1$  and  $\rho = 1$  the equation gZK is said to be Zakharov–Kuznetsov (ZK) which can be derived for Alfvén waves in a magnetized plasma at a special, critical angle to the magnetic field by means of an asymptotic multi-scale technique [8,56]. If  $b_1 = 0, b_2 = 1, b_3 = 0, b_4 = b_5 = 1$  and  $\rho = 1$ , then the equation gZK is said to be modified Zakharov–Kuznetsov (mZK) which represents an anisotropic two-dimensional generalization of the KdV equation and can be derived in a magnetized plasma for a small amplitude Alfvén wave at a critical angle to the undisturbed magnetic field. The radially symmetric positive solutions for the mZK equation have been computed [54,82].

**Example 15** Let's consider the Sharma–Tasso–Olver equation (STO) with its fission and fusion [76]

$$u_t + 3\alpha u_x^2 + 3\alpha u^2 u_x + 3\alpha u u_{xx} + \alpha u_{xxx} = 0.$$

Attention has been focused on the STO equation in [50,68] and the references therein due to its appearance in scientific applications [74].

**Example 16** Consider the Cahn–Hilliard equation

$$u_t + u_{xxxx} = (u^3 - u)_{xx} + \beta u_x.$$

The Cahn–Hilliard equation is related to a number of interesting physical phenomena like spinodal decomposition, phase separation and phase-ordering dynamics. On the other hand this equation is very hard and difficult to solve. In this paper by considering a modified extended tanh method, we found some exact solutions of the Cahn–Hilliard equation. This equation is very crucial in materials science [10,11,25]. Many articles have focused on mathematical and numerical studies of this equation [1,19,49].

## Future Directions

Nonlinear phenomena play a crucial role in applied mathematics and physics. Furthermore, when an original nonlinear equation is directly calculated, the solution will preserve the actual physical characters of solutions. Explicit solutions to the nonlinear equations are of fundamental importance. Various effective methods have been developed to understand the mechanisms of these physical

models, to help physicians and engineers and to ensure knowledge for physical problems and its applications.

Many explicit exact methods have been introduced in the literature [14,15,16,31,46,69,70,71,72,73,75]. These include the Bäcklund transformation, Hopf–Cole transformation, Generalized Miura Transformation, the Inverse scattering method, Darboux transformation, Painlevé method, homogeneous balance method, similarity reduction method, tanh method, Exp-function method, sine-cosine method and so on. There are also many numerical methods implemented for these equations [22,23,28,32,33,40,41,42,43,44,45,59,64,66]. These include the Finite elements method, finite difference methods and some approximate methods such as the Adomian decomposition method, Homotopy perturbation method, Variational perturbation method, Sinc-Galerkin method and so on.

There is still much work to be done by researchers in this field involving applications of nonlinear equations and exact and numerical implementations.

## Bibliography

### Primary Literature

1. Barrett JW, Blowey JF (1999) Finite element approximation of the Cahn–Hilliard equation with concentration dependent mobility. *Math Comput* 68:487–517
2. Benjamin TB, Bona JL, Mahony JJ (1972) Model Equations for Waves in Nonlinear Dispersive Systems. *Phil Trans Royal Soc London* 227:47–78
3. Bepalov VI, Talanov VI (1966) Filamentary structure of light beams in nonlinear liquids. *JETP Lett* 3:307–310
4. Bona JL, Pritchard WG, Scott LR (1981) An Evaluation for Water Waves. *Phil Trans Royal Soc London A* 302:457–510
5. Bona JL, Pritchard WG, Scott LR (1983) A Comparison of Solutions of two Model Equations for Long Waves. In: Lebovitz NR (ed), *Fluid Dynamics in Astrophysics and Geophysics. Lectures in Applied Mathematics*. Am Math Soc 20:235–267
6. Bona JL, Sachs RL (1988) Global Existence of Smooth Solutions and Stability Theory of Solitary Waves for a Generalized Boussinesq Equation. *Commun Math Phys* 118:15–29
7. Boussinesq J (1871) *Théorie de l'intumescence Liquid Appelée Onde Solitaire ou de Translation, se Propageant dans un Canal Rectangulaire*. *Comptes Rendus Acad Sci (Paris)* 72:755–759
8. Bridges TJ, Reich S (2001) Multi-symplectic spectral discretizations for the Zakharov–Kuznetsov and Shallow water equations. *Physica D* 152:491–504
9. Burgers J (1948) A mathematical model illustrating the theory of turbulence, *Advances in Applied Mechanics*. Academic Press, New York, pp 171–199
10. Chan JW (1961) On spinodal decomposition. *Acta Metall* 9:795
11. Chan JW, Hilliard JE (1958) Free energy of a nonuniform system I. Interfacial free energy. *J Chem Phys* 28:258–267
12. Clarkson PA, LeVeque RJ, Saxton R (1986) Solitary Wave Interactions in Elastic Rods. *Stud Appl Math* 75:95–122

13. Cole JD (1951) On a quasilinear parabolic equation occurring in aerodynamic. *Quart Appl Math* 9:225–236
14. Debnath L (1983) *Nonlinear Waves*. Cambridge University Press, Cambridge
15. Debnath L (1997) *Nonlinear Partial Differential Equations for Scientist and Engineers*. Birkhauser, Boston
16. Drazin PG, Johnson RS (1989) *Solutions: An Introduction*. Cambridge University Press, Cambridge
17. Edmundson DE, Enns RH (1992) Bistable light bullets. *Opt Lett* 17:586
18. Fisher RA (1937) The wave of advance of advantageous genes. *Ann Eugenics* 7:353–369
19. Garcke H (2000) Habilitation Thesis, Bonn University, Bonn
20. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. *Phys Rev Lett* 19:1095–1097
21. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equation and generalizations. IV. Method for exact solution. *Commun Pure Appl Math XXVII*:97–133
22. Geyikli T, Kaya D (2005) An application for a Modified KdV equation by the decomposition method and finite element method. *Appl Math Comp* 169:971–981
23. Geyikli T, Kaya D (2005) Comparison of the solutions obtained by B-spline FEM and ADM of KdV equation. *Appl Math Comp* 169:146–156
24. Grad H, Hu PN (1967) Unified shock profile in plasma. *Phys Fluids* 10:2596–2601
25. Gurtin M (1996) Generalized Ginzburg–Landau and Cahn–Hilliard equations based on a microforce balance. *Physica D* 92:178–192
26. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulse in dispersive dielectric fibers, I: Anomalous dispersion. *Appl Phys Lett* 23:142–144
27. Hayata K, Koshiba M (1995) Algebraic solitary-wave solutions of a nonlinear Schrödinger equation. *Phys Rev E* 51:1499
28. Helal MA, Mehanna MS (2007) A comparative study between two different methods for solving the general Korteweg–de Vries equation. *Chaos Solitons Fractals* 33:725–739
29. Hunter JK, Scheurle J (1988) Existence of perturbed solitary wave solutions to a model equation for water-waves. *Physica D* 32:253–268
30. Ichikawa VH (1979) Topic on solitons in plasma. *Physica Scripta* 20:296–305
31. Inan IE, Kaya D (2006) Some Exact Solutions to the Potential Kadomtsev–Petviashvili Equation. *Phys Lett A* 355:314–318
32. Inan IE, Kaya D (2006) Some exact solutions to the potential Kadomtsev–Petviashvili equation. *Phys Lett A* 355:314–318
33. Inan IE, Kaya D (2007) Exact solutions of the some nonlinear partial differential equations. *Physica A* 381:104–115
34. Jonson RS (1970) A nonlinear equation incorporating damping and dispersion. *J Phys Mech* 42:49–60
35. Jonson RS (1972) Shallow water waves in a viscous fluid-the undular bore. *Phys Fluids* 15:1693–1699
36. Kadomtsev BB, Petviashvili VI (1970) On the Stability of Solitary Waves in Weakly Dispersive Media. *Sov Phys Dokl* 15:539–541
37. Kakutani T, Ona H (1969) Weak nonlinear hydromagnetic waves in a cod collision – free plasma. *J Phys Soc Japan* 26:1305–1319
38. Karpman VI, Krushkal EM (1969) Modulated waves in nonlinear dispersive media. *Sov Phys JETP* 28:277–281
39. Kawahara TJ (1972) Oscillatory Solitary Waves in Dispersive Media. *Phys Soc Japan* 33:260
40. Kaya D (2003) A Numerical Solution of the Sine-Gordon Equation Using the Modified Decomposition Method. *Appl Math Comp* 143:309–317
41. Kaya D (2006) The exact and numerical solitary-wave solutions for generalized modified boussinesq equation. *Phys Lett A* 348:244–250
42. Kaya D, Al-Khaled K (2007) A numerical comparison of a Kawahara equation. *Phys Lett A* 363:433–439
43. Kaya D, El-Sayed SM (2003) An Application of the Decomposition Method for the Generalized KdV and RLW Equations. *Chaos Solitons Fractals* 17:869–877
44. Kaya D, El-Sayed SM (2003) Numerical soliton-like solutions of the potential Kadomtsev–Petviashvili equation by the decomposition method. *Phys Lett A* 320:192–199
45. Kaya D, El-Sayed SM (2003) On a Generalized Fifth Order KdV Equations. *Phys Lett A* 310:44–51
46. Khater AH, El-Kalaawy OH, Helal MA (1997) Two new classes of exact solutions for the KdV equation via Bäcklund transformations. *Chaos Solitons Fractals* 8:1901–1909
47. Korteweg DJ, de Vries H (1895) On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves. *Philosophical Magazine* 39:422–443
48. Li B, Chen Y, Zhang H (2003) Exact travelling wave solutions for a generalized Zakharov–Kuznetsov equation. *Appl Math Comput* 146:653–666
49. Li D, Zhong C (1998) Global attractor for the Cahn–Hilliard system with fast growing nonlinearity. *J Differ Equ* 149(2):191
50. Lian Z, Lou SY (2005) Symmetries and exact solutions of the Sharma–Tasso–Olver equation. *Nonlinear Anal* 63:1167–1177
51. Edmundson D, Enns R (1996) Light Bullet Home Page. <http://www.sfu.ca/~renns/bullets.html>
52. Ma WX (1993) An exact solution to two-dimensional Korteweg–de Vries–Burgers equation. *J Phys A* 26:17–20
53. Parkas EJ (1994) Exact solutions to the two-dimensional Korteweg–de Vries–Burgers equation. *J Phys A* 27:497–501
54. Parkes J, Munro S (1999) The derivation of a modified Zakharov–Kuznetsov equation and the stability of its solutions. *J Plasma Phys* 62:305–317
55. Pawlik M, Rowlands G (1975) The propagation of solitary waves in piezoelectric semiconductors. *J Phys C* 8:1189–1204
56. Pelinovsky DE, Grimshaw RHJ (1996) An asymptotic approach to solitary wave instability and critical collapse in long-wave KdV-type evolution equations. *Physica D* 98:139–155
57. Peregrine DH (1967) Long Waves on a Beach. *J Fluid Mech* 27:815–827
58. Peregrine DH (1996) Calculations of the Development of an Undular Bore. *J Fluid Mech* 25:321–330
59. Polat N, Kaya D, Tutalar HI (2006) A analytic and numerical solution to a modified Kawahara equation and a convergence analysis of the method. *Appl Math Comp* 179:466–472
60. Rayleigh L (1876) *On Waves*. The London and Edinburgh and Dublin Philosophical Magazine 5:257
61. Robert WM, Vsvolod VA, Yuri SK, John DL (1996) Optical solitons with power-law asymptotics. *Phys Rev E* 54:2936
62. Russell JS (1844) Report on Waves. 14th meeting of the British Association for the Advancement of Science. BAAS, London
63. Schimizu K, Ichikawa VH (1972) Auto modulation of ion oscillation modes in plasma. *J Phys Soc Japan* 33:789–792
64. Shawagfeh N, Kaya D (2004) Series solution to the Pochhammer–Chree equation and comparison with exact solutions. *Comp Math Appl* 47:1915–1920

65. Su CH, Gardner CS (1969) Derivation of the Korteweg–de Vries and Burgers' equation. *J Math Phys* 10:536–539
66. Ugurlu Y, Kaya D, Solution of the Cahn–Hilliard equation. *Comput Math Appl* (accepted for publication)
67. Wang M, Li LX, Zhang J (2007) Various exact solutions of nonlinear Schrödinger equation with two nonlinear terms. *Chaos Solitons Fract* 31:594–601
68. Wang S, Tang X, Lou SY (2004) Soliton fission and fusion: Burgers equation and Sharma–Tasso–Olver equation. *Chaos Solitons Fractals* 21:231–239
69. Wazwaz AM (2002) *Partial Differential Equations: Methods and Applications*. Balkema, Rotterdam
70. Wazwaz AM (2007) Analytic study for fifth-order KdV-type equations with arbitrary power nonlinearities. *Comm Nonlinear Sci Num Sim* 12:904–909
71. Wazwaz AM (2007) A variable separated ODE method for solving the triple sine-Gordon and the triple sinh-Gordon equations. *Chaos Solitons Fractals* 33:703–710
72. Wazwaz AM (2007) The extended tanh method for abundant solitary wave solutions of nonlinear wave equations. *Appl Math Comp* 187:1131–1142
73. Wazwaz AM, Helal MA (2004) Variants of the generalized fifth-order KdV equation with compact and noncompact structures. *Chaos Solitons Fractals* 21:579–589
74. Wazwaz AM (2007) New solitons and kinks solutions to the Sharma–Tasso–Olver equation. *Appl Math Comp* 188:1205–1213
75. Whitham GB (1974) *Linear and Nonlinear Waves*. Wiley, New York
76. Yan Z (2003) Integrability for two types of the  $(2 + 1)$ -dimensional generalized Sharma–Tasso–Olver integro-differential equations. *MM Res* 22:302–324
77. Zabusky NJ (1967) *Nonlinear Partial Differential Equations*. Academic Press, New York
78. Zabusky NJ (1967) A synergetic approach to problems of nonlinear dispersive wave propagations and interaction. In: Ames WF (ed) *Proc. Symp. on Nonlinear Partial Differential equations*. Academic Press, Boston, pp 223–258
79. Zabusky NJ, Kruskal MD (1965) Interactions of solitons in a collisionless plasma and the recurrence of initial states. *Phys Rev Lett* 15:240–243
80. Zhang W, Chang Q, Fan E (2003) Methods of judging shape of solitary wave and solution formulae for some evolution equations with nonlinear terms of high order. *J Math Anal Appl* 287:1–18
81. Zhou CT, He XT, Chen SG (1992) Basic dynamic properties of the high-order nonlinear Schrödinger equation. *Phys Rev A* 46:2277
82. Munro S, Parker EJ (1997) The stability of solitary-wave Solutions to a modified Zakharov–Kuznetsov equation. *J Plasma Phys* 64:411–426

## Books and Reviews

The following, referenced by the end of the paper, is intended to give some useful for further reading.

For another obtaining of the KdV equation for water waves, see Kevorkian and Cole (1981); one can see the work of the Johnson (1972) for a different water-wave application with variable depth, for waves on arbitrary shears in the work of Freeman and Johnson (1970) and Johnson (1980) for a review of one and

two-dimensional KdV equations. In addition to these; one can see the book of Drazin and Johnson (1989) for some numerical solutions of nonlinear evolution equations. In the work of the Zabusky, Kruskal and Deam (F1965) and Eilbeck (F1981), one can see the motion pictures of soliton interactions. See a comparison of the KdV equation with water wave experiments in Hammack and Segur (1974)

For further reading of the classical exact solutions of the nonlinear equations can be seen in the works: the Lax approach is described in Lax (1968); Calogero and Degasperis (1982, A.20), the Hirota's bilinear approach is developed in Matsuno (1984), the Bäcklund transformations are described in Rogers and Shadwick (1982); Lamb (1980, Chap. 8), the Painleve properties is discussed by Ablowitz and Segur (1981, Sect. 3.8), In the book of Dodd, Eilbeck, Gibbon and Morris (1982, Chap. 10) can found review of the many numerical methods to solve nonlinear evolution equations and shown many of their solutions.

---

## Patterns and Interfaces in Dissipative Dynamics

L.M. PISMEN

Department of Chemical Engineering and Minerva  
Center for Nonlinear Physics of Complex Systems,  
Technion – Israel Institute of Technology, Haifa, Israel

## Article Outline

[Glossary](#)  
[Definition of the Subject](#)  
[Introduction](#)  
[Stationary Patterns](#)  
[Moving Interfaces](#)  
[Wave Patterns](#)  
[Future Directions](#)  
[Bibliography](#)

## Glossary

**Interface** An interface separates domains where different stationary states or different patterns prevail. In the latter case, it is also called a domain wall. The interface typically has a finite thickness comparable to a characteristic intrinsic scale of the system but small compared to the overall system size.

**Stationary pattern** A stationary pattern is formed as a result of an instability to perturbations with a finite wavenumber. It may have any of various spatial structures (striped, square, hexagonal, or quasicrystalline in 2D, lamellar, crystalline or quasicrystalline in 3D) and may slowly evolve in time.

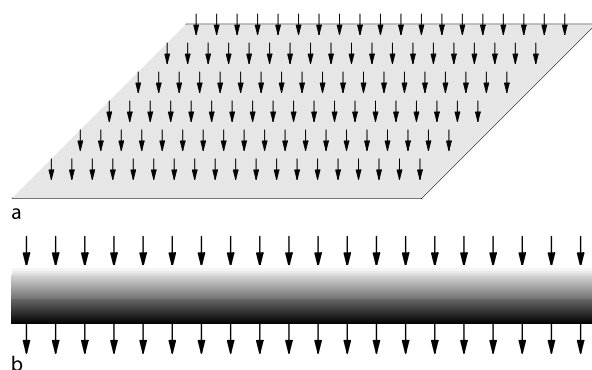
**Wave pattern** A wave pattern is formed by a combination of waves propagating in one or different directions.

### Definition of the Subject

A *pattern* is an inhomogeneous state of a physical system that arises spontaneously under spatially homogeneous conditions. Spontaneous pattern formation has been first observed by Faraday [1] in vibrated liquid layers and Bénard [2] in fluids heated from below. Turing [3] envisaged pattern formation as the mechanism of morphogenesis in living Nature. Some patterns can be described as a collection of patches or domains where one of alternative homogeneous states prevails, separated by relatively narrow *interfaces*. In their turn, moving interfaces may develop corrugation patterns. Patterns can be stationary or wavelike; they can be regular, interlaced by defects, or chaotic (turbulent). In the latter part of 20th century, numerous pattern formation phenomena have been observed in chemistry, biology, fluid mechanics, granular media, nonlinear optics, and other applications, and common models describing these phenomena in physically dissimilar settings have been formulated and studied. Understanding pattern formation is important both for describing natural self-organization phenomena and for developing manufacturing processes based on self-organization.

### Introduction

A typical setup of a non-equilibrium system that may undergo a symmetry-breaking transition is shown in Fig. 1. A non-equilibrium stationary state homogeneous in the “horizontal” plane is sustained by fluxes in the normal (“vertical”) direction, along which an inhomogeneous



**Patterns and Interfaces in Dissipative Dynamics, Figure 1**

An open system isotropic in two dimensions. A truly two-dimensional system (*above*) and a cut through a system with vertical structure (*below*, shown symbolically by *varied shading*). Arrows indicate the direction of external fluxes

“vertical structure” may be formed. This setup may be realized as a layer of fluid or granular matter; a chemically reacting system, such as an active layer or a catalytic surface; an area where different populations spread out and compete; a propagating interphase boundary, e.g. a melting or crystallizing solid; a slice of nonlinear optical medium, etc. Under certain conditions, most commonly, under increased driving, this homogeneous state may be destabilized, giving way to a stationary or moving pattern with a characteristic wavelength dependent on physical properties of the system as well as on external fluxes. In chemically reacting systems, three-dimensional patterns can be also formed when a sufficient amount of reactants is stored; such patterns may exist, of course, for a limited time until the original cache is depleted. Mathematically, a pattern typically emerges as an inhomogeneous solution of a (system of) partial differential equation(s) with space-independent coefficients in the absence of lateral fluxes.

Alternative states, corresponding to different phases, may exist also in equilibrium systems. Following a fast quench past a critical point, different states, separated by domain boundaries, would be approached at spatially removed locations. Typically, these domains would consequently slowly coarsen to minimize the extent of an interphase boundary and related energetic costs. A stationary pattern with a finite wavelength may exist, however, also at equilibrium, provided it minimizes the free energy of the system. Such patterns are realized as “mesoscopic crystals” in block-copolymers consisting of two kinds of mutually repelling units [4].

In fluid mechanics, inhomogeneous states, most often disordered but still retaining a measure of regularity, are commonplace, as anybody observing wavy sea and cloud patterns could have realized long before classical 19th century experiments of Faraday and Bénard. Wave patterns generated by oscillatory chemical reactions (which long considered to be impossible due to thermodynamic misconceptions) were demonstrated in 1960s [5], while controlled experiments demonstrating persistent stationary chemical patterns in reaction-diffusion systems had to wait till early 1990s [6]. Shell growth patterns [7], striped and dotted animal skins [8], and desert vegetation patterns [9] have been always here for anybody to observe, before finding rational explanation in terms of the same nonlinear models. Corrugated interfaces were observed and described both as flame fronts in combustion theory [10] and as dendrite forms of growing crystals [11]. More recently, much attention has been drawn by nonlinear optical patterns – spontaneous images emerging in optical circuits and lasers [12].



## Stationary Patterns

### Symmetry-Breaking Transitions

The most direct way to formation of stationary patterns is a symmetry-breaking bifurcation. It can be demonstrated in a straightforward way taking as an example a two-component reaction-diffusion system (RDS)

$$\partial_t u = D_1 \nabla^2 u + \gamma_1^{-1} f(u, v), \quad (1)$$

$$\partial_t v = D_2 \nabla^2 v + \gamma_2^{-1} g(u, v), \quad (2)$$

where  $f(u, v)$ ,  $g(u, v)$  are source functions depending on the variables  $u$  and  $v$ ,  $D_1$ ,  $D_2$  are diffusivities, and  $\nabla^2$  is the Laplace operator. We suppose that the system has a homogeneous stationary state (HSS)  $u = u_s$ ,  $v = v_s$  satisfying  $f(u_s, v_s) = g(u_s, v_s) = 0$ ; the factors  $\gamma_1$ ,  $\gamma_2$  are introduced to scale the derivatives  $f_u$ ,  $g_v$  computed at this HSS to unity. Stability analysis of the chosen HSS to infinitesimal perturbations  $\tilde{u}, \tilde{v} \propto \exp(i \mathbf{k} \cdot \mathbf{x})$  with a wave vector  $\mathbf{k}$  shows that the most dangerous perturbations have the wavenumber

$$|\mathbf{k}|^2 \equiv k^2 = \frac{1}{2} \left( \frac{f_u}{\gamma_1 D_1} + \frac{g_v}{\gamma_2 D_2} \right). \quad (3)$$

This value should be positive, which is possible only in the presence of positive feedback, or, in chemical terms, when at least one of the species is “autocatalytic”, say,  $f_u > 0$ . Breaking of spatial symmetry preempts Hopf bifurcation, which occurs at  $\gamma_1^{-1} f_u + \gamma_2^{-1} g_v = 0$  and leads to homogeneous oscillations, provided only one of the species is autocatalytic, so that  $g_v < 0$ , and the autocatalytic species is less diffusive. Thus, for spatial symmetry breaking in a two-component system, one needs a combination of a slowly diffusing “activator” and a rapidly diffusing “inhibitor”.

The development of a pattern can be understood qualitatively in the following way. A local upsurge of the activator concentration increases also the concentration of the inhibitor, which spreads out suppressing the activator at neighboring locations. This, in turn, suppresses the inhibitor locally and, through inhibitor diffusion, enhances the activator further along the line, so that the inhomogeneous state spreads out. This scheme works with the roles of an activator and an inhibitor played, respectively, by prey and predator in population dynamics, by growing plants and seeping moisture in ecology, or, rather less directly, by buoyancy and heat conduction in natural convection.

Pattern formation may also result from nonlocal interactions. For example, a nonlocal extension of the nonlin-

ear Schrödinger equation (NLS) for a complex field  $u$ ,

$$-i \partial_t u = \nabla^2 u - u(x) \int U(\mathbf{x} - \boldsymbol{\xi}) |u(\boldsymbol{\xi})|^2 d\boldsymbol{\xi}, \quad (4)$$

generates a patterned state known as “supersolid”, as compared and contrasted to superfluid solutions of the local NLS [13]. It might be possible to derive nonlocal equations from a local RDS. Thus, if in Eq. (2)  $\gamma_2 \ll \gamma_1$ , so that the inhibitor is fast as well as diffusive, the time derivative can be neglected; then, if the function  $g(u, v)$  is linear in  $v$ , Eq. (2) can be resolved with the help of an appropriate Green’s function, and substituting it in Eq. (1) yields a nonlocal activator equation.

### Selection of Stationary Patterns

Symmetry breaking transitions in more than one dimension are degenerate due to spatial symmetries. In an isotropic system, an arbitrary number of differently directed modes with  $k = |\mathbf{k}| = \text{idem}$  can be excited beyond the bifurcation point. A combination of these modes can give a variety of distinct *planforms*. Competition among the modes that determines the pattern selection is described by *amplitude equations* describing evolution of complex amplitudes  $a_j$ , which have a general form

$$\begin{aligned} \frac{da_j}{dt} &= -\frac{\partial V}{\partial \bar{a}_j}, \\ V &= -\mu \sum |a_j|^2 + \sum v_{ijk} a_i a_j a_k \\ &\quad + \sum v_{ijkl} a_i a_j a_k a_l + \text{c.c.} \end{aligned} \quad (5)$$

Here the coefficient  $\mu$  is proportional to the deviation from the bifurcation point; real coefficients  $v_{ijk}$ ,  $v_{ijkl}$  characterize nonlinear interactions among the modes; the summation is carried out over all closed polygons formed by the wave vectors of extant modes. The product of the amplitudes  $\bar{a}_j$ ,  $\bar{a}_k$ , etc. (where the overline denotes the complex conjugate) may appear in the equation for the amplitude  $a_i$  if the respective wave vectors add up to zero,  $\mathbf{k}_i + \mathbf{k}_j + \mathbf{k}_k + \dots = 0$ . This condition ensures that the modes in question are in resonance. Otherwise, the product of these modes rapidly oscillates and is averaged out when the amplitude equation is derived using a multiscale expansion procedure. Stationary solutions, i.e. potential minima of Eq. (5) with one, two, three, or more non-vanishing modes with a symmetric star of wave vectors correspond, respectively, to a striped, square, hexagonal, or quasicrystalline pattern.

The cubic term in the potential (5) generates the lowest-order, hence, strongest nonlinear interactions. This



term vanishes in the presence of inversion symmetry  $a \rightarrow -a$ , which exists, in particular, in the thoroughly studied case of buoyancy-driven convection in the Boussinesq approximation. Otherwise, it is dominant near the bifurcation point, causing (in 2D) a subcritical transition to a hexagonal pattern comprising modes forming a regular triangle. These three modes are in resonance, which means that their phases are not independent but bound by a linear relationship. The sum of phases always adjusts in such a way that interactions are destabilizing. The remaining two phase degrees of freedom correspond to translational symmetry in the plane.

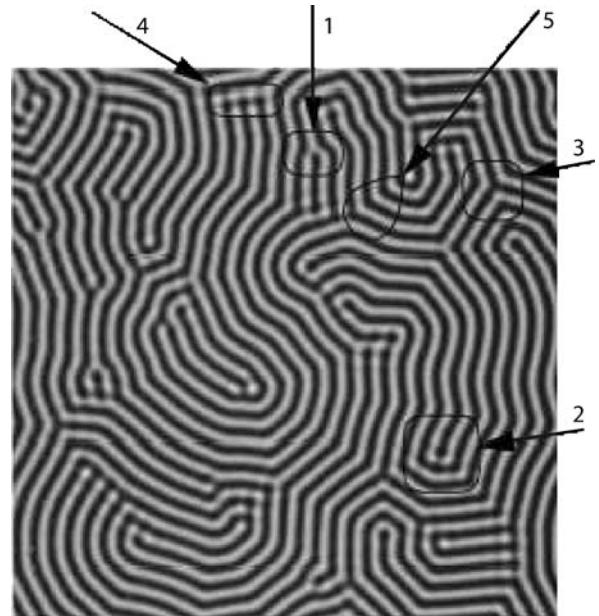
In 3D, the preferred patterns, or crystalline structures, comprise wave vectors forming a regular polyhedron with triangular faces – tetrahedron, octahedron or dodecahedron [14]. The former two correspond to a body-centered cubic (bcc), and the last one, to a quasicrystalline structure with fivefold symmetry. These lowest-order interactions cannot, however, stabilize the pattern at a finite amplitude, and next-order interactions generated by the quartic term in Eq. (5) are necessary to saturate the pattern. Depending on respective interaction coefficients, various structures can be chosen.

A greater variety of patterns may arise if planforms with different wavenumbers  $k$  are excited simultaneously. This can be achieved in a most natural way in two-layer systems where the wavelength of the excited pattern depends on the thickness of each layer, as in convection [15], or different diffusivities, as in a pattern-forming chemical system [16]. More possibilities arise in nonlinear optics where spatial symmetry breaking may occur on different wavelengths at rather close values of a control parameter [17]. The resulting coupled amplitude equations can generate a variety of composite planforms, which may have a form of superstructures or quasicrystals. Lowest-order interactions can generate various resonances; no rigid fitting of wavenumbers is required for this, since resonant modes can form an isosceles triangle. Dynamics of mode interactions may be complicated [18], since the gradient structure of Eq. (5) is, generally, lost.

Regular patterns may suffer various instabilities, which limit the range of admissible wavelengths or lead to a change of the planform through excitation of a non-collinear mode or decay of an extant mode. Wavelength changing instabilities, as a rule, do not saturate and lead to formation of defects.

### Modulated and Distorted Patterns

Natural patterns seen both in experiment and simulations are never perfect: their amplitudes may be modulated at



Patterns and Interfaces in Dissipative Dynamics, Figure 2

Various forms of pattern defects. 1 – dislocation, 2 – concave disclination, 3 – convex disclination, 4 – amplitude domain wall, 5 – phase domain wall ([19], reproduced with permission. Copyright by the American Physical Society)

distances large compared to the basic wavelength, and they may have various defects: dislocations, disclinations, and domain walls. An example of an imperfect striped pattern is shown in Fig. 2. Variation of local wavelengths is possible because instability spreads out to a finite range of wavenumbers, scaled as the square root of the parametric deviation from the bifurcation point. Other imperfections are a consequence of the rotational symmetry of the system. Different orientations of stripes may be chosen at different locations, either randomly or under influence of boundary conditions or local inhomogeneities. The discrepancies of local orientations are reconciled through formation of disclinations and domain walls, while dislocations reconcile discrepancies of local wavelengths.

Weak distortions, which do not contain defects, can be described by means of either space-dependent amplitude equations applicable to small-amplitude patterns near the bifurcation point, or phase dynamics applicable also to finite-amplitude patterns but restricted to long-scale distortions.

The amplitude equation must have an anisotropic form in an isotropic system, the source of anisotropy being the direction of the wave vector itself. Modulations of this amplitude along and across the direction of the wave vector  $\mathbf{k}$  should be scaled differently, since adding a small lon-

gitudinal component, say,  $\epsilon q_x$  changes  $k = |\mathbf{k}|$  by  $O(\epsilon)$ , while adding a transverse component of the same magnitude  $\epsilon q_y$  changes  $k$  by  $O(\epsilon^2)$  only; thus the stripes are bent far more easily than they are compressed or extended. This leads to the Newell–Whitehead–Segel (NWS) amplitude equation [20,21], which can be written in a rescaled universal form

$$\partial_t u = \left( \partial_x - \frac{i}{2k} \partial_y^2 \right)^2 u + u - |u|^2 u. \quad (6)$$

The mixed-order differential operator entering this equation precisely accounts for the equivalence of all structures with identical wavenumbers, independently of the direction of the wave vector.

The NWS equation is ill-suited for computations, since the orientation of the coordinate axes depends on the local phase gradient, so that the differential operator is in fact strongly nonlinear. Most model computations of striped patterns are based on the Swift–Hohenberg (SH) equation

$$\partial_t u = -(1 + \nabla^2)^2 u + u(\mu - u^2). \quad (7)$$

In an *anisotropic* system where a certain direction of stripes is preferred, the situation is easier, and the amplitude equation can be reduced by rescaling to an *isotropic* real Ginzburg–Landau (RGL) equation

$$\partial_t u = \nabla^2 u + u - |u|^2 u. \quad (8)$$

### Phase Dynamics

The idea of phase dynamics [22] is to characterize a striped pattern by means of a single variable – phase  $\theta$ , which changes by  $2\pi$  over the period of the pattern or, more conveniently, by a rescaled phase  $\Theta = \epsilon\theta$ . The derivatives of the phase are the wave vector  $\mathbf{k} = \nabla\theta$  and frequency  $\omega = -\theta_t$ , which vary on an extended scale exceeding the wavelength of the underlying structure by a factor  $\epsilon^{-1} \gg 1$ . The general form of the phase equation in an

isotropic system is determined by scaling and symmetry considerations alone:

$$\partial_T \Theta = D_1 (\mathbf{n} \cdot \hat{\nabla})^2 \Theta + D_2 \hat{\nabla}^2 \Theta, \quad (9)$$

where  $\partial_T, \hat{\nabla}$  are derivatives with respect to slow time and extended spatial variables,  $\mathbf{n}$  is the unit vector along  $\mathbf{k}$ , and  $D_1, D_2$  are phase diffusivities that depend on a particular underlying problem and are, generally, functions of  $k$ . This equation can be also presented in an elegant gradient form [23].

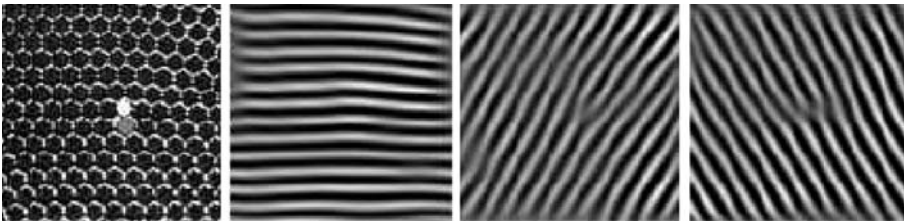
The phase equation (9) is, in fact, strongly nonlinear due to the dependence of both the diffusivities and the direction of the unit vector  $\mathbf{n}$  on the local phase gradient. It can be linearized, yielding an anisotropic diffusion equation, only when deviations from a prevailing wave vector  $\mathbf{k} = \mathbf{k}_0$  are arbitrary small. If the  $X$ - and  $Y$ -axes are drawn, respectively, along and across  $\mathbf{k}_0$ , (9) reduces to

$$\Theta_T = D_{\parallel}(k_0) \Theta_{XX} + D_{\perp}(k_0) \Theta_{YY}, \quad (10)$$

where  $D_{\parallel} = D_1 + D_2$  and  $D_{\perp} = D_2$  are, respectively, the longitudinal and transverse phase diffusivities. The pattern with the wavenumber  $k_0$  is stable to long-scale perturbations when both phase diffusivities are positive. Vanishing  $D_{\parallel}$  corresponds to the *Eckhaus* instability and vanishing  $D_{\perp}$  to the *zigzag* instability. Eckhaus instability defines the upper limit of stable wavenumbers. It never saturates, and usually leads to formation of defects effectively increasing the wavelength. Zigzag instability defines the upper limit of stable wavenumbers; it causes bending of stripes effectively decreasing the wavelength.

### Dynamics of Defects

Dynamics of strongly distorted patterns is mostly governed by motion and interaction of defects. Defects are topological objects [24]: a dislocation is characterized by circulation of the phase around any enclosing contour equal to an integer multiple of  $2\pi$ , and a disclination, by



Patterns and Interfaces in Dissipative Dynamics, Figure 3

Hexagonal pattern containing a penta-hepta defect (left) and its three constituent modes obtained by Fourier filtering of the initial image ([32], reproduced with permission)

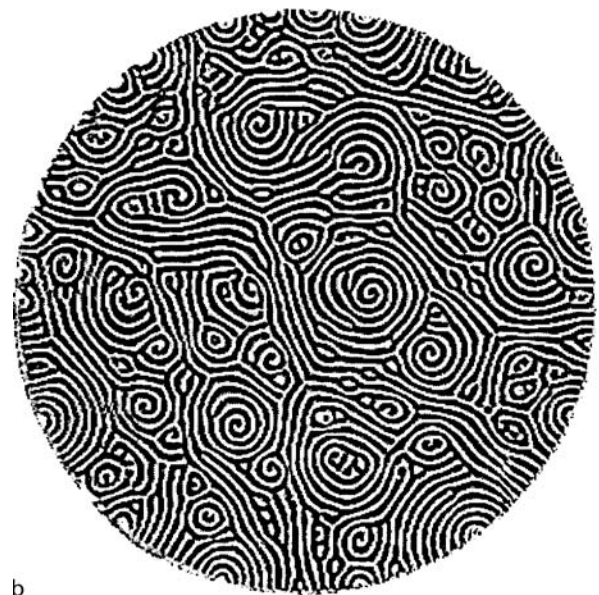
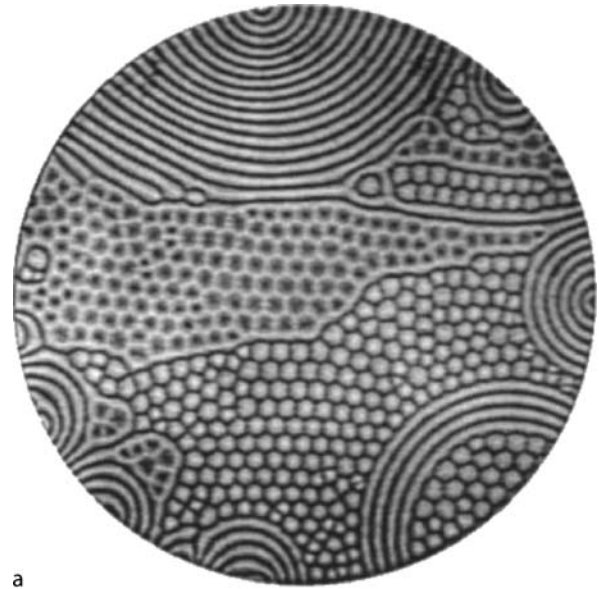
circulation of the direction of the wave vector equal to an integer multiple of  $\pi$ . A single dislocation climbing across the direction of the wave vector of a striped pattern effects a change of the wavenumber over an extended region. The force driving the dislocation is due to the deviation from the optimal wavenumber. Eckhaus instability of a striped pattern leads to the formation of a dislocation pair. It is notable that, although the far field of dislocations can be described by phase equations, their interaction is determined by the dislocation core where these equations are inapplicable [25,26].

Motion of dislocations in striped patterns is well understood and supported by experimental evidence [27] for anisotropic patterns governed by Eq. (8). The structure of dislocations in isotropic systems described by Eq. (6) is more complicated, being strongly anisotropic [28]. Disclinations pose more difficulties for the analysis, even on the topological level [24], see [29].

Paradoxically, defects enhance relaxation of the pattern to a state of minimum energy corresponding to an “optimal” wavelength. If a deviation of the control parameter from the symmetry breaking bifurcation point is of  $O(\epsilon^2)$ , the width of the band of excited modes is of  $O(\epsilon)$ , but the band width actually observed in a natural patterns containing defects is of  $O(\epsilon^2)$  [19]. The band shrinks due to motion of point defects and adjustments influenced by domain walls.

The structure and interaction of dislocations in a hexagonal pattern is strongly affected by the resonant character of interactions among the constituent modes. Dislocations in any two modes of the triplet forming a hexagonal pattern, created originally at arbitrary locations, are always attracted to each other [30,31], eventually forming an immobile bound pair corresponding to a penta-hepta defect (see Fig. 3).

Equations (6), (8) are derivable from an energy functional that decreases monotonically in time until a stationary state of minimal energy is reached; this state may still contain defects necessary to satisfy boundary conditions in a confined region. In some cases, however, an additional field, besides the amplitude, is necessary to adequately describe a physical system even close to the symmetry-breaking bifurcation point. A well known example is Bénard convection in low Prandtl number fluids where the additional factor is mean flow generated by pattern distortions and advecting the entire pattern. In this case, the patterns remains weakly turbulent indefinitely long, displaying labyrinthine structures, coexisting striped and hexagonal domains [33] or spiral defect chaos [30] (see Fig. 4). Chaotic non-stationary patterns also typically appear at higher amplitudes. In reaction-diffusion systems non-sta-



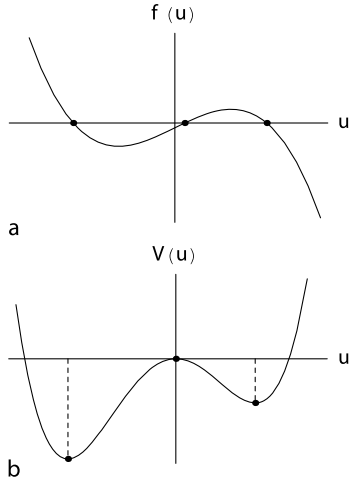
**Patterns and Interfaces in Dissipative Dynamics, Figure 4**  
**a** Coexisting domains; **b** Spiral defect chaos [30] (reproduced with permission)

tionary and chaotic patterns become more likely when the inhibitor response is slowed down.

## Moving Interfaces

### Stationary and Propagating Fronts

Many physical systems, either at equilibrium or in a non-equilibrium steady state sustained by external fluxes, may exist in two or more alternative states. If different states are



**Patterns and Interfaces in Dissipative Dynamics, Figure 5**  
A function  $f(u)$  with three zeros (a) and the respective double-well potential (b)

attained at different spatial locations, they are separated by an *interface*, carrying excess energy. The simplest model is a single “reaction–diffusion” equation

$$\partial_t u = D \nabla^2 u + f(u), \quad (11)$$

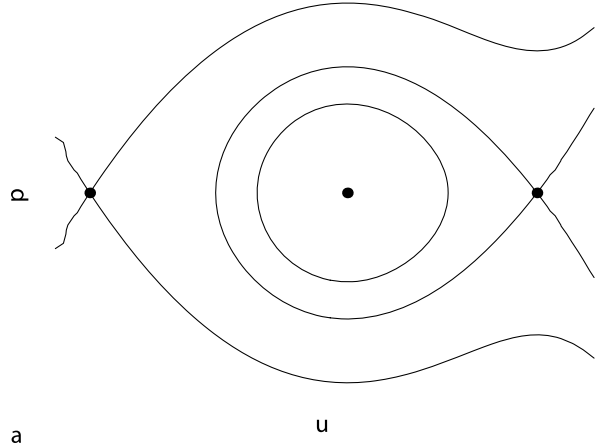
where  $D$  is diffusivity and the function  $f(u) = -V'(u)$  (see Fig. 5) has three zeroes that correspond to two stable (with  $f'(u) < 0$ ) and one unstable (with  $f'(u) > 0$ ) HSS. This equation was first used in the context of phase equilibria [34] as a model of gas–liquid interface, with  $u$  denoting density. It was later extended to the solidification problem, with  $u$  denoting a fictitious “phase field” assuming its two stable values  $u = u_s^\pm$  in the liquid and solid phases [35]. The coefficient  $D$  is interpreted in this context as *rigidity*. The “reaction–diffusion” interpretation applies to non-equilibrium systems, such as a catalytic surface or an ecological domain, with  $u$  denoting concentration and  $f(u)$ , the net production rate.

A straight-line or planar interface is stationary when the potentials  $V(u_s^\pm)$  are equal. It carries then the interfacial energy

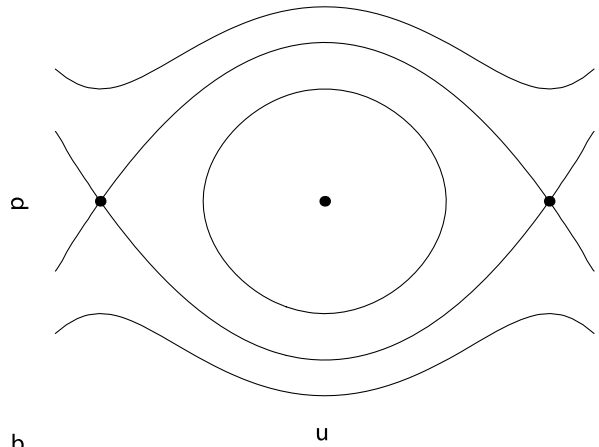
$$\sigma = D \int_{-\infty}^{\infty} u'(x)^2 dx = \int_{u_s^-}^{u_s^+} \sqrt{2DV(u)} du, \quad (12)$$

which is identified with surface tension.

If the potentials are unequal, the front moves in the direction decreasing the total energy of the system. Assuming that the motion is stationary and directed along the  $x$  axis, (11) can be rewritten in the comoving frame propagating with the front velocity  $c$ . The steadily propagating solution depends on a single coordinate  $\xi = x - ct$ , and



a



b

**Patterns and Interfaces in Dissipative Dynamics, Figure 6**  
Generic trajectories in the phase plane  $u, p = u'(x)$  (a) and a non-generic set of trajectories containing a heteroclinic orbit (b)

(11) reduces to an ordinary differential equation

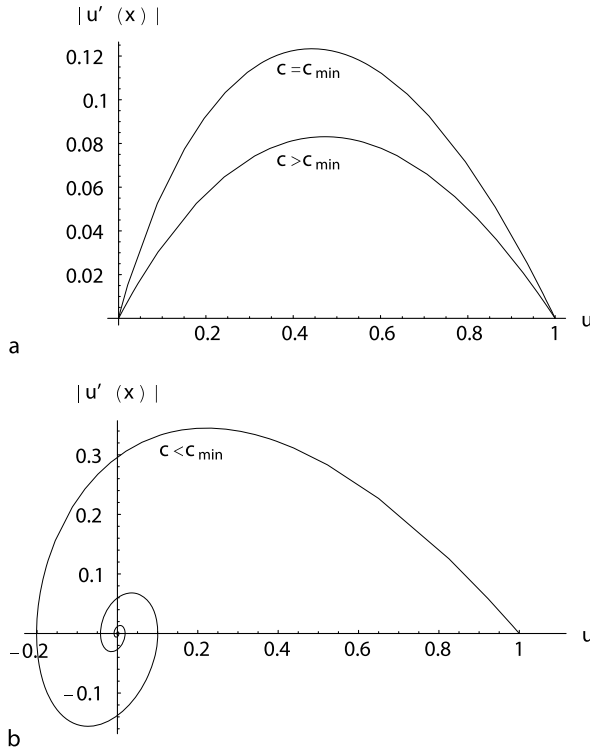
$$cu'(\xi) + Du''(\xi) + f(u) = 0, \quad (13)$$

subject to the boundary conditions  $u = u_s^\pm$  at  $\xi \rightarrow \pm\infty$ . When both equilibria are stable, they are saddles when viewed as equilibria of (13). The front solution corresponds to a heteroclinic trajectory connecting the equilibria  $u = u_s^\pm$ . The heteroclinic connection exists only at unique value of  $c$  (see Fig. 6); thus, the propagation speed is determined uniquely by solving a nonlinear eigenvalue problem. Its value is proportional to the difference of potentials of the two HSS:

$$c = \frac{D}{\sigma} \Delta V, \quad \Delta V = V(u_s^-) - V(u_s^+). \quad (14)$$

The situation is different when the retreating state  $u = u_s^0$  is unstable. This often happens in population dy-



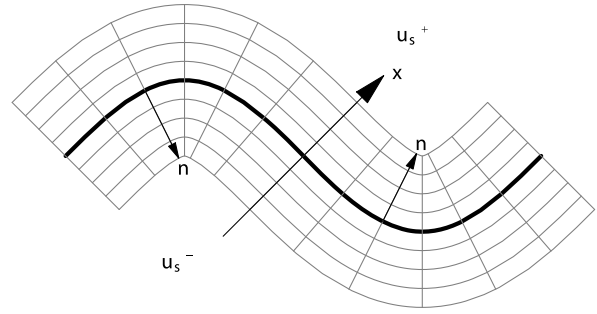


**Patterns and Interfaces in Dissipative Dynamics, Figure 7**  
Trajectories in the phase plane connecting a stable and an unstable equilibrium

namics: a state where a competitively advantageous species is absent is formally unstable to infinitesimal perturbations but will be nevertheless preserved at any location until this species is introduced there. An unstable state, viewed as an equilibrium point of (13), is a stable node at propagation speeds exceeding a certain threshold  $c_{\min}$ . Thus, a trajectory starting from the advancing stable HSS connects generically to  $u_s^0$  at any  $c > c_{\min}$  (see Fig. 7). Actual propagation speed is selected dynamically at the leading edge [37,38], and turns out to be equal to the minimum speed  $c_{\min}$ , which corresponds to the steepest front profile. Under certain conditions (when overshoots are allowed) a faster speed corresponding to a still steeper profile is selected nonlinearly [38]. In the former case, the front is “pulled” by perturbations growing at the leading edge and described by linearized equations, while in the latter case, it is “pushed” by nonlinear interactions favoring the advancing state.

### Interfacial Instabilities

The front solution is neutrally stable to translations along the  $x$ -axis. This neutral (Goldstone) mode is weakly per-



**Patterns and Interfaces in Dissipative Dynamics, Figure 8**  
Construction of the aligned coordinate frame. The coordinate lines are shown in gray. Arrows show the local directions of the normal  $n$  and the  $x$ -axis. Observe a singularity developing on the concave side

turbed when the translation is weakly nonuniform, so that the front becomes curvilinear but the curvature radius still far exceeds the characteristic front thickness.

Propagation of a weakly curved front is best understood in a coordinate frame aligned with its deformed shape. The nominal front position is defined by replacing a diffuse transitional region by a planar curve  $C$  drawn along some intermediate level of the variable  $u$ . The coordinate lines  $x = \text{const}$  are obtained by shifting the curve along the normal by a constant increment, as shown in Fig. 8. This shift causes the length to increase on convex, and to decrease on concave side of the curve. Eventually, a singularity develops in the latter direction, but, when the curvature radius is much larger than the characteristic front thickness, this will happen far away within the region where one of the HSS is approached.

When (13) is rewritten in the aligned frame an expanded viewing the curvature as a small parameter, the local normal propagation speed of a curved front is expressed by the eikonal equation

$$c = c_0 - D\kappa = \frac{D}{\sigma} (\Delta V - \sigma\kappa), \quad (15)$$

where  $c_0$  is the speed of a planar front and  $\kappa$  is the Gaussian curvature.

Since convex front segments propagate slower and concave segments faster, the front tends to flatten, provided  $c_0$  is uniform everywhere. Instabilities may arise, however, when  $c_0$  increases ahead of the front. This may happen in the presence of an externally imposed gradient, as in directional solidification [11], but most commonly is caused by an additional “control” field. The control field responsible for the Mullins–Sekerka instability of solidification fronts [11,39] is the concentration of a contaminant, which is rejected by the solid and slows down so-



lidification by lowering the melting temperature. Since the contaminant diffuses away more easily from convex segments, they tend to propagate faster, which causes instability when the driving is strong enough to overcome surface tension.

Another example is instability of a combustion front, which separates hot burnt-out and cold fuel-rich domains [10]. A thin front structure arises in this case because combustion requires both fuel and sufficient temperature for its initiation, and both fuel concentration and temperature play the role of control variables. When heat transfer is the limiting factor, convex segments cool down and propagate slower, and the front is stable. When, on the opposite, propagation is limited by fuel supply, convex segments accelerate and instability sets on, leading to corrugated fronts.

Dynamics of weak deviations  $\zeta(\mathbf{y})$  from a stable planar front spanned by a 2-vector  $\mathbf{y}$  is described by expanding the normal propagation speed, front curvature and the control field in powers of a small parameter scaling both the deviation  $\zeta$  and its transverse derivative  $\nabla_{\mathbf{y}}$ , as well as time. For stable fronts, the appropriate scaling is  $\zeta = O(1)$ ,  $\nabla_{\mathbf{y}} = O(\epsilon)$ ,  $\partial_t = O(\epsilon^2)$ , leading to the Burgers equation

$$\partial_t \zeta = D \nabla_{\mathbf{y}}^2 \zeta - \frac{1}{2} c_0 |\nabla_{\mathbf{y}} \zeta|^2. \quad (16)$$

The particular coefficients here correspond to (15), but also in other cases the same universal form can be obtained after the coefficients are removed by rescaling, provided the effective diffusivity  $D$  is positive. If the latter is negative but small,  $|D| = O(\epsilon^2)$ , the appropriate scaling is  $\zeta = O(\epsilon)$ ,  $\nabla_{\mathbf{y}} = O(\epsilon)$ ,  $\partial_t = O(\epsilon^4)$ , and expanding to a higher order yields, after scaling away the coefficients, the Kuramoto–Sivashinsky equation [40]

$$\partial_t \zeta + \nabla_{\mathbf{y}}^2 \zeta + (\nabla_{\mathbf{y}}^2)^2 \zeta + \frac{1}{2} |\nabla_{\mathbf{y}} \zeta|^2 = 0. \quad (17)$$

This equation, appearing also in phase dynamics [41], is a paradigm of weak turbulence.

### Front Interactions and Coarsening

Fronts of opposite polarity in a one-dimensional system attract and eventually coalesce, thereby coarsening the distribution of domains, which may have been created initially in the process of phase separation or relaxation to alternative HSS. The interaction is, however, very weak, falling off exponentially with separation. In higher dimensions, the principal cause of coarsening, or Ostwald ripening, is the curvature dependence of the propagation speed, whereby small droplets with high curvature tend to shrink and eventually disappear. This is a manifestation of the

Gibbs–Thomson effect relating the equilibrium conditions with the radius of a droplet.

Coarsening most often occurs under conditions when evolution is constrained by a conservation law, so that the integral  $\int u(\mathbf{x}) d\mathbf{x}$  expressing the total amount of material in the system remains constant. Under these conditions, fronts cannot move independently from each other. The conservation law is accounted for when (11) is replaced by the Cahn–Hilliard equation [35]

$$\partial_t u = \nabla^2 \mu, \quad \mu = -[D \nabla^2 u + f(u)]. \quad (18)$$

The eikonal equation governing the front motion retains the form (15), but the value  $c_0$  depends on chemical potential  $\mu$ . The latter shifts in the course of coarsening in such a way that the value of the critical radius  $R = \kappa^{-1}$  of a droplet that neither grows or shrinks, keeps growing as smaller droplets disappear. Analytical theory [42] predicts universal asymptotic droplet size distribution at late stages of coarsening.

### Structures Built up of Fronts

Coarsening can be precluded when changes in an additional control field arrest growth of large and shrinking of small domains. This leads to formation of a variety of patterns and solitary structures. The paradigmatic system for exploring these phenomena is the FitzHugh–Nagumo system, which has the form (1), (2) with the function  $f(u, v)$  cubic in  $u$  and linear in  $v$  and a linear function  $g(u, v)$ . The rescaled form suitable for the analysis of stationary structures is

$$\epsilon^2 \partial_t u = \epsilon^2 \nabla^2 u + u - u^3 - \epsilon v, \quad (19)$$

$$\tau^{-1} \partial_t v = \nabla^2 v - v - v + \mu u, \quad (20)$$

Here  $\epsilon = \sqrt{\gamma_1 D_1 / \gamma_2 D_2} \ll 1$  is the ratio of the characteristic lengths associated with the activator and the inhibitor,  $\tau = D_2 / D_1$ ; the small coupling parameter  $\epsilon$  in (19) ensures a balance between the effect of small interfacial curvature and weak symmetry breaking between the alternative HSS  $u_s^\pm = \pm 1 + O(\epsilon)$ ; the remaining parameters  $\mu$  and  $v$  regulate the coupling stress and bias.

Structures generated by the system (19), (20) are built up by assigning a region where the activator approaches one of the alternative HSS, computing the respective inhibitor distribution, and finding stationarity conditions for the fronts forming the boundaries of this region [43]. Possible stationary structures in two dimensions are a solitary band, a solitary disk, a striped pattern, or a hexagonal grid consisting of almost circular spots. The size of spots

or stripes is determined by the parameters of the system, but there is a considerable leeway in choosing the general configuration. Under certain conditions, it even might be possible to store information by creating or extinguishing spots at chosen locations [44]. In other cases, splitting of a solitary spot initiates a multiplication cascade [45], leading eventually to a hexagonal pattern filling the plane.

Instabilities of stationary structures are studied with the help of the linearized eikonal equation (15) combined with the inhibitor equation (20) where the last term is expressed through a shift of the front position. Both solitary bands and disks can suffer zigzag (leading eventually to splitting), oscillatory and traveling instabilities. The latter two become prevalent as the parameter  $\tau$  decreases, so that the inhibitor response to front displacements slows down. For example, a solitary band is destabilized in the zigzag mode at  $\tau > 1$ , while the traveling instability comes first at smaller  $\tau$  (see Fig. 9). Oscillatory instability is always preceded by traveling one in this case, but may become relevant for a solitary disk.

Traveling instability indicates transition to various propagating structures and wave patterns. A solitary spot tends to either dissolve or spread out sidewise after being immobilized; in the latter case, a spiral structure starts to develop as the ends lag behind. A traveling spot can be, however, stabilized if a second inhibitor, both fast and long-range, is added [46].

Various patterns of propagating fronts can be generated beyond this limit by the same FitzHugh–Nagumo system, which, however, should be scaled differently for this purpose. Unlike stationary or slowly evolving patterns where the characteristic length scale is set by the diffu-

sional range of the long-scale inhibitor, the wavelength of a propagating pattern is tied to the propagation speed and remains finite even when the inhibitor is nondiffusive. The long scale should be redefined therefore on the basis of the characteristic propagation speed of the activator front  $c^* = \sqrt{D_1/\gamma_1}$  and the characteristic relaxation time of the inhibitor  $\gamma_2$ . Using this “advective” length unit,  $L^* = \gamma_2 \sqrt{D_1/\gamma_1}$  brings (1), (2) to the dimensionless form

$$\gamma \partial_t u = \gamma^2 \nabla^2 u + f(u, v), \quad (21)$$

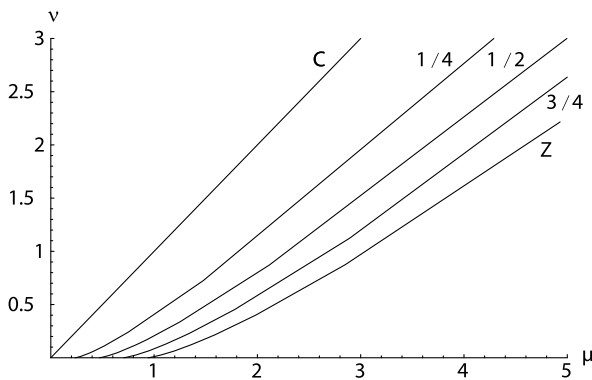
$$\partial_t v = \delta^2 \nabla^2 v + g(u, v), \quad (22)$$

where  $\gamma = \gamma_1/\gamma_2$ ,  $\delta = \gamma/\epsilon = \sqrt{\gamma D_2/D_1}$ . The “inner” scale of the transitional layer, where the system switches between the two alternative activator states,  $u = u_s^\pm$ , is now set exclusively by the capacitance ratio  $\gamma$ , independently of diffusivities, and, provided  $\gamma \ll 1$ , remains small even when the inhibitor is less diffusive than the activator. The parameters can be chosen in such a way that  $\delta \ll 1$ , so that the inhibitor diffusion is negligible, provided  $\gamma \ll D_1/D_2$ . Under these conditions, the inhibitor diffusion can be neglected, reducing (22) to  $\partial_t v = g(u, v)$ . Although this equation contains no mechanism for healing discontinuities in  $v$ , the inhibitor field should remain smooth in the course of evolution, barring freaky initial conditions or strongly localized perturbations. This opens the easiest way of constructing various wave patterns, including such exotic objects as chaotic wave trains [47].

## Interfaces of Patterns

Interfaces between different patterns or different pattern orientations (domain walls) can be described in the simplest way on the level of amplitude equations. This may give qualitatively correct results in static problems, even though changes across a domain wall in patterns generated in simulations and experiments are usually effected on a length comparable with the prevailing wavelength of the pattern. One can expect that a stationary solution exists only when the wavelengths are equal on both sides of the wall; otherwise, the wall would propagate in the direction decreasing the overall energy of the pattern. It turns out that an even stronger restriction is true, and both wavelengths should be optimal [48]. In this way, domain walls, alongside dislocations, enhance relaxation of the pattern to the optimal wavelength.

Dynamic problems are strongly influenced by detailed structure of the pattern, which is lost on the level of amplitude equations. When a pattern advances into an unstable



**Patterns and Interfaces in Dissipative Dynamics, Figure 9**  
Existence boundary (C) and loci of zigzag (Z) and traveling instability for a solitary band. The loci of traveling instability are marked by respective values of  $\tau$ . A stable band exists between the line C and an applicable instability locus

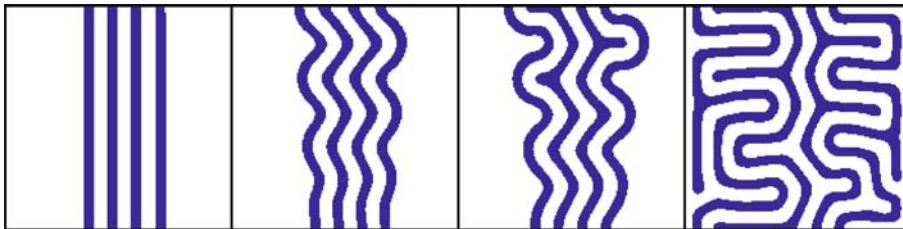


Patterns and Interfaces in Dissipative Dynamics, Figure 10

A scheme of depinning transitions showing crystallization (C) and melting (M) thresholds for an infinite cluster, as well as the corresponding limits for clusters of different sizes, terminating in single-cell limits 1-C, 1-M

uniform state, the wavelength selected at the leading edge is not identical to the wavelength of the full-grown pattern formed behind the front, and neither one coincides with the optimal wavelength [49].

In the case when a stable homogeneous solution coexists with a stable periodic pattern, stable stationary fronts between the two states exist within a finite parametric interval [50], rather than at a single point where the energies of both states are equal, as amplitude equations would predict. The motion of this front is affected by the discrete structure of the pattern, which causes self-induced pinning hindering the retreat of a metastable state. There are two depinning transitions, corresponding to “crystallization” or “melting” of the pattern, shown schematically by thick lines in Fig. 10. Between the two limits, various metastable stationary structures exist: a single cell (“soliton”), a finite patterned inclusion, sandwiched between semi-infinite domains occupied by a uniform state, or a semi-infinite pattern, coexisting with a uniform state. To the right of the crystallization threshold C, the pattern advances by a periodic nucleation process which creates new elementary cells at the interface [51], while to the left of the melting limit M, the pattern recedes as elementary cells at the interface are destroyed. A different, far more efficient depinning mechanism works in two dimensions [52]. It is initiated by a zigzag instability of the pattern followed by nucleation of disclinations, which further move toward the uniform state, as seen in Fig. 11. This generates stripes extending in the normal direction, turning eventually the original boundary into a domain wall separating striped patterns rotated by  $\pi/2$ .



Patterns and Interfaces in Dissipative Dynamics, Figure 11

Depinning of striped pattern initiated by a zigzag instability [52]

## Wave Patterns

### Plane Waves

A simplest propagating wave pattern is a periodic solution depending on a moving coordinate  $\xi = x - ct$ , where  $c = \omega/k$  is phase velocity,  $\omega$  is frequency and  $k$  is wavenumber. A waveform  $\sim \exp[i(kx - \omega t)]$  may emerge directly by symmetry breaking bifurcation, but this is not the most common mechanism. It is impossible, in particular, in a two-component RDS (1), (2), where other scenarios lead to wave patterns. One of them, mentioned in the preceding Section, is traveling instability of stationary structures. Another road to wave patterns, most amenable to analytical tools, starts in the vicinity of a Hopf bifurcation, where small-amplitude oscillations weakly modulated in space are described by the complex Ginzburg–Landau (CGL) equation. Its standard rescaled form is

$$\partial_t u = (1 + i\eta)\nabla^2 u + u - (1 + i\nu)|u|^2 u. \quad (23)$$

A plane wave solution of (23) with the wave vector  $\mathbf{k}$  is

$$u = \rho_0 \exp[i(\mathbf{k} \cdot \mathbf{x} - \omega t)], \quad (24)$$

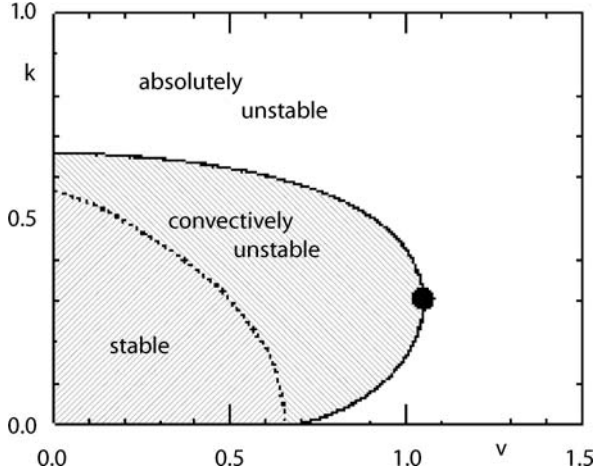
$$\rho_0 = \sqrt{1 - k^2}, \quad \omega = \nu + (\eta - \nu)k^2.$$

The waves are dispersive, and the group velocity is  $\mathbf{v} = 2\mathbf{k}(\eta - \nu)$ .

Instabilities of plane waves are studied most efficiently with the help of the phase dynamics approach, since the most dangerous perturbation modes can be viewed as long-scale distortions of neutrally stable translational modes. The longitudinal and transverse phase diffusivities are

$$D_{\parallel} = \frac{1 + \nu\eta - k^2(3 + \nu\eta + 2\nu^2)}{1 - k^2}, \quad D_{\perp} = 1 + \nu\eta. \quad (25)$$

Vanishing  $D_{\parallel}$  marks the threshold of Eckhaus instability, which limits the range of stable wavenumbers. Vanishing  $D_{\perp}$  signals Benjamin–Feir (self-focusing) instability, independently of the wavelength. Both instabilities arising



**Patterns and Interfaces in Dissipative Dynamics, Figure 12**  
Limits of convective and absolute instabilities in the plane  $(v, k)$  for  $\eta = -3/2$ . The dot marks the limit of convectively unstable waves ([53], reproduced with permission. Copyright by the American Physical Society)

at the respective thresholds are convective, which means that growing perturbations are washed away with the prevailing group velocity. The absolute instability condition stipulating growth of perturbation at a particular location is less restrictive (see Fig. 12). Numerical simulations [53] show that transition to turbulence occurs only when the absolute stability condition is violated, but the system is very sensitive to noise in the convectively unstable region.

Besides uniform wave trains, there is a variety of non-uniform one-dimensional solutions of the CGL equation with a constant frequency and spatially varying modulus and wavenumber, which are stationary in a frame propagating with a certain speed  $c$  and depend on the comoving coordinate  $\xi = x - ct$  only. The solutions approaching asymptotically at  $\xi \rightarrow \pm\infty$  either plane waves or the trivial state can be also viewed as defects separating domains where different uniform states prevail. Such solutions include pulses, approaching the trivial state at both extremes; nonlinear fronts, separating the trivial state from an invading wave train, and domain boundaries separating plane waves directed in the opposite sense and, possibly, having different wavelength [53,54]. Interactions among various defects dominate chaotic dynamics beyond the self-focusing instability limit [55].

Amplitude equations for wave patterns emerging directly from an HSS through a symmetry breaking bifurcation with  $\omega \neq 0$ ,  $k \neq 0$  should account for competition between waves with amplitudes  $u^\pm$  propagating in the opposite directions, which may either suppress one another

or combine to a standing wave. The normalized form of coupled equations for  $u^\pm$  is

$$\partial_t u^\pm \pm c u_x^\pm = (1 + i\eta)u_{xx}^\pm + u^\pm - (1 + i\nu_+) |u^\pm|^2 u^\pm - g(1 + i\nu_-) |u^\mp|^2 u^\pm, \quad (26)$$

where  $g$  is a coupling parameter. The orders of magnitude of all terms of these equations can be balanced only when the phase velocity  $c = \omega/k$  is of the same  $O(\epsilon)$  as  $u^\pm$ . Generically,  $c = O(1)$ , and the advective term  $c u_x^\pm$  is dominant. For a single wave, it can be removed by transforming to the comoving frame. When both waves are present, each wave, viewed in its own frame  $\xi_\pm = x \mp ct$  samples the average amplitude of its counterpart propagating in this frame with a fast speed. The appropriate amplitude equations have then the form [56]

$$\partial_t u^\pm = (1 + i\eta)u_{\xi_\pm \xi_\pm}^\pm + u^\pm - (1 + i\nu_+) |u^\pm|^2 u^\pm - g(1 + i\nu_-) |u^\mp|^2 u^\pm. \quad (27)$$

These equations retain only global coupling carried by the spatial averages  $\langle |u^\mp|^2 \rangle$ .

In two dimensions, the amplitude equations also involve resonant interactions of pairs of waves propagating in the opposite directions. This makes possible complex dynamics even when the amplitudes are uniform and obey space-independent equations [57]

$$\partial_t u_1^\pm = u_1^\pm [\mu - \nu_+ |u_1^\pm|^2 - \nu_- |u_1^\mp|^2 - \beta (|u_2^\pm|^2 + |u_2^\mp|^2)] + \gamma \bar{u}_1^\mp u_2^\pm u_2^\mp. \quad (28)$$

### Spiral and Scroll Waves

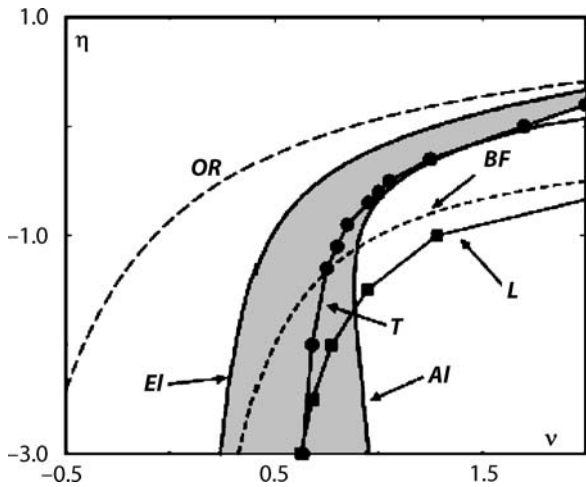
A ubiquitous and extensively studied waveform is a rotating spiral wave. Its specific feature is the presence of a phase singularity. An  $n$ -armed spiral wave can be constructed as a circularly symmetric vortex solution of (23) with the topological charge  $n$ , i. e. phase circulation  $2\pi n$ . Unlike a symmetric defect in (8), the phase must also depend on the radial coordinate, so that the vortex radiates a wave with a certain uniquely selected asymptotic wavenumber  $k_\infty$ . This solution is obtained [58] in polar coordinates  $r, \phi$  by assuming an *ansatz*

$$u = \rho(r) e^{i\theta}, \quad \theta = n\phi + \psi(r) - \omega t. \quad (29)$$

Using this *ansatz* brings (23) to the form

$$\rho''(r) + r^{-1} \rho'(r) + (1 - k^2 - n^2/r^2 - \rho^2) \rho = 0, \quad (30)$$

$$\frac{1}{\rho^2} \frac{d}{dr} (rk\rho^2) = q(\rho_\infty^2 - \rho^2), \quad (31)$$

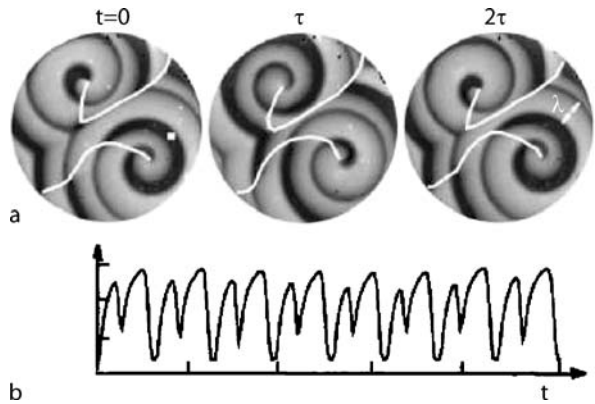


Patterns and Interfaces in Dissipative Dynamics, Figure 13

Stability limits of a spiral wave solution in the parametric plane ( $\eta$ ,  $\nu$ ). The curve *EI* shows the limit of convective instability and *AI*, of absolute instability for the waves emitted by the spiral; *OR* is the boundary of the oscillatory spatial decay for the emitted waves,  $q = 0.845$  (bound states exist to the right of this line). *BF* indicates the Benjamin-Feir limit  $\nu\eta = -1$ , *L* is the limit of phase turbulence, and *T* corresponds to the transition to defect turbulence for random initial conditions ([53], based on [59]; reproduced with permission). (Copyright by the American Physical Society)

where  $k = \psi'(r)$  is the radial wavenumber. Stability analysis of plane waves applies also to far regions of spiral waves; one could expect therefore a transition to a turbulent state to occur under conditions when the selected asymptotic wavenumber  $k_\infty$  falls into the range where the corresponding plane wave solution of (23) is unstable. The respective stability limits in the parametric plane ( $\eta$ ,  $\nu$ ) are presented in Fig. 13.

Another approach to constructing rotating spiral waves exploits kinematics of fronts of opposite polarity described by RDS (21), (22) [60]. The inhibitor diffusion can be neglected almost everywhere, except in the crucial tip region where the two fronts meet. Behavior of the spiral tip and its meandering instability has been elucidated analytically using a multiscale technique matching different approximations in overlapping regions [61]. Complex dynamics of a meandering tip, which exhibit quasiperiodic and chaotic motion in some parametric domains, can be well described with the help of a simpler phenomenological model [62]. A similar instability of spiral waves described by the CGL equation is the core acceleration instability [53], which may serve as a trigger of transition to spatio-temporal chaos alternative to instability of radiated waves.



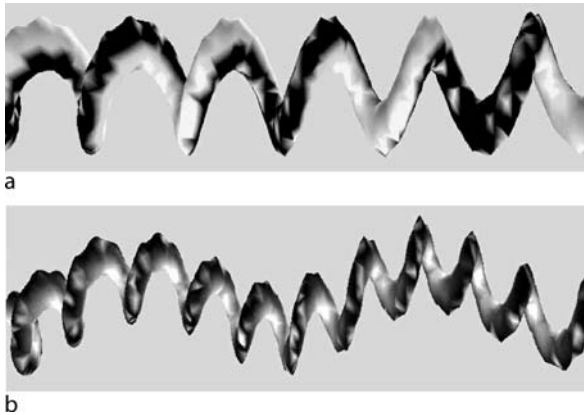
Patterns and Interfaces in Dissipative Dynamics, Figure 14

a A pair of period two spiral waves with the fundamental period  $\tau$  and the average wavelength  $\lambda$ . The white solid lines are the synchronization defects. b A period two time series measured at the point marked by the white filled square ([63], reproduced with permission. Copyright by the American Physical Society)

A special kind of spiral wave patterns arises when the underlying dynamical system undergoes a period doubling transition. The period doubling causes the appearance of synchronization defect (SD) lines, which serve to reconcile the doubling of the oscillation period with the period of rotation of the spiral wave (see Fig. 14a). These lines are defined as the loci of those points in the medium where the two loops of the period two orbit exchange their positions in local phase space. The period two oscillations on the opposite sides of a SD are shifted relative to each other by  $2\pi$  (i. e., a half of the full period), so that the dynamics projected on the rotation direction is effectively of period one, while it is of period two locally at any point in the medium (Fig. 14b).

A three-dimensional extension of a rotating spiral is a rotating scroll wave. The core filament of a scroll wave is a line vortex. A scroll wave with a straight-line core directed along the  $z$ -axis has identical spiral waves in each cross-section. Even then, the structure can be nontrivial if the spiral phases are given a phase twist, i. e. are shifted along the  $z$ -axis. A curved core filament may also close up into a ring or even form knots. A stable scroll structure evolves to decrease the filament curvature [64]. This kind of dynamics is similar to curvature-driven motion of interfaces, but may be reversed when the filament is unstable. The most dangerous perturbation modes are long-scale modes associated with meandering or translational core deformations [65]. Meandering instability usually saturates as a distorted scroll wave with a twisted rotating core (Fig. 15). Instability in the translation mode,





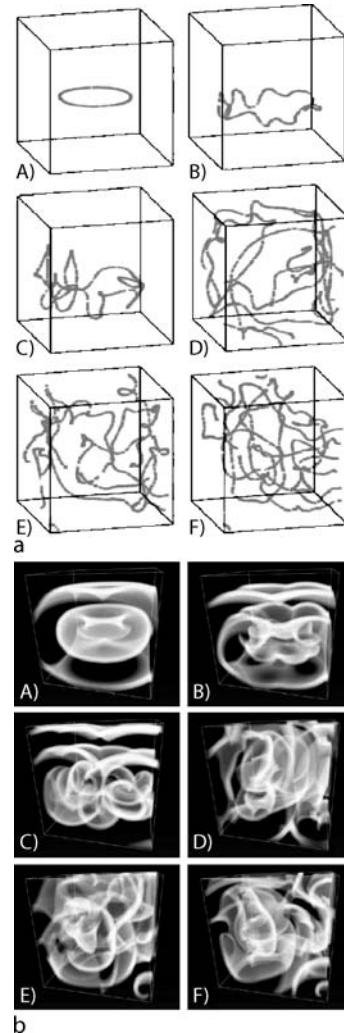
**Patterns and Interfaces in Dissipative Dynamics, Figure 15**

**a** A restabilized helical vortex; **b** A doubly periodic “superhelix” Isosurfaces of the modulus  $\rho = 0.6$  shaded by phase field are shown. (CGL simulations [66], reproduced with permission. Copyright by the American Physical Society)

which causes spontaneous bending of the scroll axis, does not saturate, but gives rise to a scroll wave with a continuously extending core (Fig. 16a). This leads to a turbulent state visualized as a tangle of breaking wave fronts (Fig. 16b).

### Spiral Patterns and Turbulence

Interaction of spiral waves is dominated by shocks – domain boundaries where waves emanating from different centers collide. The shocks effectively screen different spiral domains from radiation emitted by other spiral cores. A typical example of a spiral domain pattern in a stable parametric range obtained in a CGL simulation run starting from random initial conditions [59] is shown in Fig. 17. At the initial stage, the system tends to relax locally to the stable state with unity real amplitude, but, as the phases are random, the relaxation is frustrated, and a large number of defects – vortices of unit charge – are formed. At the following coarsening stage, oppositely charged vortices annihilate, so that the density of defects decreases. The coarsening process, however, stops halfway, leaving a certain number of single-charged spiral vortices with either sense of rotation. Vortices that failed to conquer a sufficiently large domain are reduced to “naked cores”, left to satisfy the topological condition of conservation of circulation. The resulting stable spiral domain pattern is called vortex glass. The waves always propagate outwards from the vortex cores, so that the entire domain structure is generated when local order spreads out from centers to the periphery. Perturbations, also traveling outwards with the prevailing group velocity, are absorbed at shocks, and

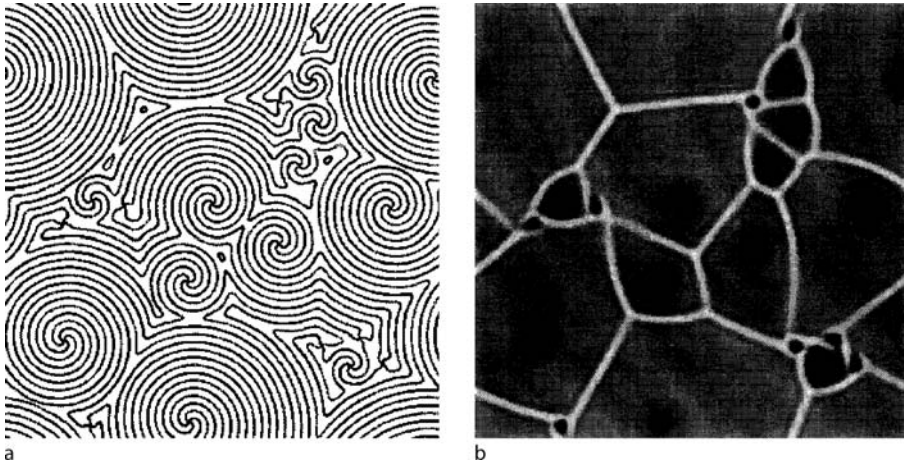


**Patterns and Interfaces in Dissipative Dynamics, Figure 16**

**Transition to turbulence due to core filament extension and breakup of scroll waves. a** Snapshots of the core filament, starting from a closed loop. **b** Respective snapshots of wave patterns showing semitransparent visualization of the activator fronts ([67], reproduced with permission. Copyright by the American Physical Society)

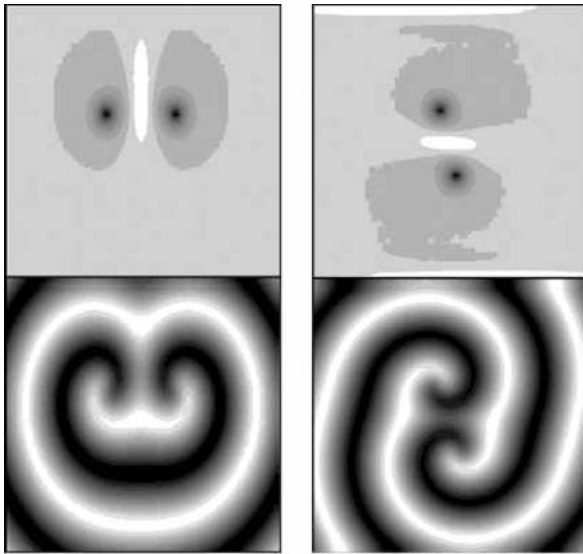
therefore the pattern may survive beyond the convective instability threshold. The turbulent state takes over only when the emanated waves become absolutely unstable, i. e., when some perturbations grow locally in the laboratory frame.

The overall structure of the pattern changes in the range of oscillatory spatial decay of waves emanated by the spiral cores (below the line OR in Fig. 13). Under these conditions, formation of stable bound spiral pairs becomes possible (see Fig. 18). Unlike the monotonic range, spiral domains may have in oscillatory range a wide size distri-



Patterns and Interfaces in Dissipative Dynamics, Figure 17

Spiral domains. *Left*: levels of constant phase. *Right*: grayscale amplitude map showing enhanced amplitudes at the shocks (CGL simulations [59], reproduced with permission from Elsevier Science)



Patterns and Interfaces in Dissipative Dynamics, Figure 18

Bound states of oppositely (*left*) and likely (*right*) charged spirals (CGL simulations,  $\eta = 0$ ,  $\nu = 1.5$ ). The images show the modulus  $\rho(x, y)$  (top) and  $\text{Re}(u)$  (bottom) ([53], reproduced with permission. Copyright by the American Physical Society)

bution, since shocks can be immobilized at different separations.

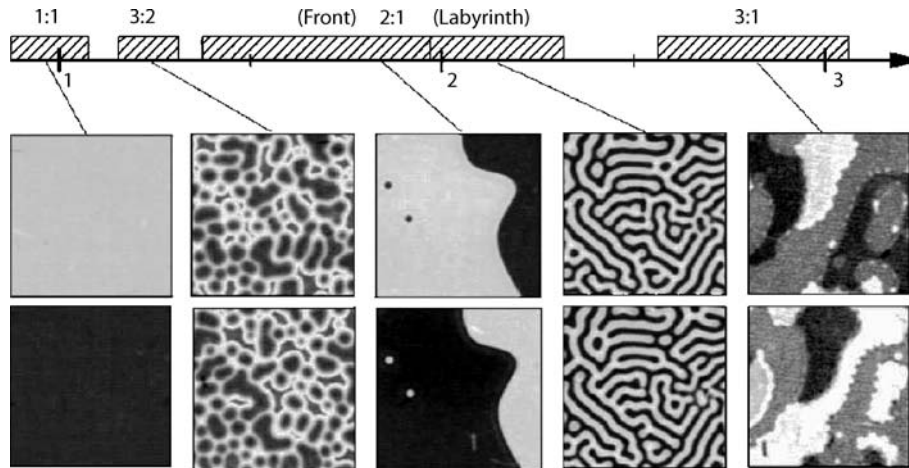
“Frozen” glassy patterns actually evolve on a very long time scale, as revealed in very long simulation runs [68]. In the monotonic range, spiral cores perform very slow diffusive motion; the apparent diffusivity increases with vortex density. In contrast, in the oscillatory range, spiral population spontaneously segregates after a very long transient into two distinct phases: large and almost immobile spi-

ral and clusters of trapped small vortices. When the “liquid fraction” is small, the resulting pattern exhibits slow intermittent dynamics: bursts of activity separated by long quiescent intervals. The system keeps evolving on an extremely slow scale, which is consistent with exponentially weak repulsion between well separated spiral cores.

Another possibility, realized in a different parametric region, is a dynamic chaotic state that shows no persistent features. This state is attained under conditions when either spiral waves or vortex cores, or both, are unstable. One can distinguish between mild phase turbulence when no phase singularities occur, and defect chaos characterized by persistent creation and annihilation of vortex pairs. Phase turbulence may persist in the parametric region between the Benjamin–Feir line and the line  $L$  in Fig. 13 [59]. Beyond the line  $L$ , defects are created spontaneously, leading to defect chaos. Transition from vortex glass to defect turbulence in simulations starting from random initial conditions occurs at the numerically determined line  $T$  in Fig. 13 [59]. The transition occurs somewhat prior to the absolute instability limit determined by the linear stability analysis of plane waves emitted by spirals. This limit can be approached, however, by starting from carefully prepared initial conditions in the form of large spirals. Prior to the transition, one can observe transient defect turbulence which is unstable to spontaneous nucleation of spirals from the “turbulent sea”, leading eventually to a vortex glass state.

### Forced Systems

External forcing, including spatially as well as temporally variable inputs, can be used in a straightforward



Patterns and Interfaces in Dissipative Dynamics, Figure 19

Frequency-locked regimes (experiment with light-sensitive reaction under periodic optical forcing [71], reproduced with permission). The axis above shows the ratio of the forcing to basic frequency. Patterns are shown in pairs, one above the other, at times separated by the forcing period  $2\pi/\omega_c$ , except for the 1 : 1 resonance where the interval is  $\pi/\omega_c$ .

way to enhance or suppress spontaneously emerging patterns [69]. Alternatively, it may enhance complexity by introducing additional spatial and temporal resonances, which may lead to formation of quasicrystalline structures [70]. Resonant forcing of oscillatory systems may drastically change the structure of wave patterns through phase locking. This happens when the CGL equation is forced on a frequency  $\omega_c$  commensurate with the basic frequency  $\omega_0$  at the Hopf bifurcation. For an integer ratio  $\omega_c/\omega_0 = n$ , the amplitude equation amending (23) can be written by adding the forcing term possessing the required symmetry:

$$\partial_t u = (1 + i\eta)\nabla^2 u + (\mu + i\omega)u - (1 + i\nu)|u|^2 u + \gamma \bar{u}^{n-1}, \quad (32)$$

where  $\gamma$  is the forcing amplitude and  $\epsilon^2\omega$  is weak effective detuning, due to both parametric deviations from the Hopf bifurcation point and weak mismatch between  $\omega_c/n$  and  $\omega_0$ . The forcing term breaks the symmetry of the CGL equation to phase rotations, reducing it to discrete symmetry  $u \rightarrow e^{i\pi m} u$ ,  $m = 1, \dots, n-1$ . This changes the character of defects: instead of vortices, one can observe fronts separating alternative phase states.

Various patterns at different forcing frequencies, which can be modeled by (32), were observed both in experiments and simulations [71,72]. Some typical patterns are shown in Fig. 19. For the case of strong resonance ( $n = 1$ ), this system provides a convenient tool for studying transitions between stationary and propagating fronts [73], labyrinthine patterns [74], and solitary structures [75]. These structures are not unlike those ob-

served in the FitzHugh–Nagumo system, although they represent standing waves with the alternative phases interchanging within each domain. Higher resonances create still more complex dynamics involving interactions of different kinds of fronts [76].

### Future Directions

The study of pattern formation is now a mature discipline based on well-established general theory and wealth of experimental evidence. The center of attention is turning to specific applications; among them, nonlinear optics and studies of granular media come to the forefront. Forcing and control of patterns, either enhancing or suppressing the complexity of behavior, are studied in detail. As a humble laptop turns into a supercomputer, more fascinating patterns, envy of abstract expressionists, are generated by model equations of increased complexity. Patterns showing dazzling mix of order and chaos are seen as well in various experimental setups. The ultimate aim of controlled creation of self-organized structures still remains elusive, and new ideas are awaited as the new century comes of age. The study of pattern formation, dealing with ubiquitous problems of order and chaos, is bound to find its way into basic curricula and wealth of practical applications.

### Bibliography

#### Primary Literature

1. Faraday M (1831) *Philos Trans R Soc Lond* 121:299
2. Bénard H (1900) *Ann Chim Phys* 7 (Ser 23) 62

3. Turing AM (1952) *Philos Trans R Soc Lond Ser B* 237:37
4. Hamley IW (2003), *Nanotechnology* 14:R39
5. Burger M, Field RJ (1985) *Oscillations and Traveling Waves in Chemical Systems*. John Wiley, New York
6. Ouyang Q, Swinney HL (1991) *Nature* 352:610
7. Gierer A, Meinhard H (1972) *Kybernetik* 12:30
8. Murray JD (1981) *J Theor Biol* 88:161
9. Gilad E, Shachak M, Meron E (2007) *Theor Popul Biol* 72:214
10. Zeldovich YB (1985) *The Mathematical Theory of Combustion and Explosions*. Consultants Bureau, New York
11. Langer JS (1980) *Rev Mod Phys* 52:1
12. Arecchi FT (1999) *Phys Rep* 318:1
13. Jossierand C, Pomeau Y, Rica S (2007) *Eur Phys J Special Topics* 146:47
14. Alexander S, McTague J (1978) *Phys Rev Lett* 41:702
15. Proctor MRE, Jones CA (1988) *J Fluid Mech* 188:301
16. Yang L, Dolnik M, Zhabotinsky AM, Epstein IR (2002) *Phys Rev Lett* 88:208303
17. Pampaloni E, Residori S, Soria S, Arecchi FT (1997) *Phys Rev Lett* 78:1042
18. Pismen LM, Rubinstein BY (1999) *Chaos Soliton Fractal* 10:761
19. Bowman C, Newell AC (1998) *Rev Mod Phys* 70:289
20. Newell AC, Whitehead JA (1969) *J Fluid Mech* 38:279
21. Segel LA (1969) *J Fluid Mech* 38:203
22. Pomeau Y, Manneville P (1979) *J Phys Lett* 40:L609
23. Cross MC, Newell AC (1984) *Physica D* 10:299
24. Mermin ND (1979) *Rev Mod Phys* 51:591
25. Bodenschatz E, Pesch W, Kramer L (1988) *Physica D* 32:135
26. Pismen LM, Rodriguez JD (1990) *Phys Rev A* 42:2471
27. Braun E, Steinberg V (1991) *Europhys Lett* 15:167
28. Nepomnyashchy AA, Pismen LM (1991) *Phys Lett A* 153:427
29. Newell AC, Passot T, Bowman C, Ercolani N, Indik R (1996) *Physica D* 97:185
30. Bodenschatz E, Pesch W, Ahlers G (2000) *Annu Rev Fluid Mech* 32:709
31. Rabinovich MI, Tsimring LS (1994) *Phys Rev E* 49:R35
32. Abou B, Wesfreid JE, Roux S (2000) *J Fluid Mech* 416:217
33. Assenheimer M, Steinberg V (1993) *Phys Rev Lett* 70:3888
34. van der Waals JD (1894) *Z Phys Chem* 13:657
35. Cahn JW, Hilliard JE (1958) *J Chem Phys* 28:258
36. Sagués F, Sancho JM, García-Ojalvo J (2007) *Spatiotemporal order out of noise*. *Rev Mod Phys* 79:829
37. Kolmogorov A, Petrovsky I, Piskunov N (1937) *Bull Univ Moscow, Ser Int Sec A* 1:1
38. van Saarloos W (2003) *Phys Rep* 386:29
39. Mullins WW, Sekerka RF (1963) *J Appl Phys* 34:323
40. Sivashinsky GI (1977) *Acta Astronaut* 4:1177
41. Kuramoto Y, Tsuzuki T (1976) *Prog Theor Phys* 55:356
42. Lifshitz IM, Slyozov VV (1958) *Zh Eksp Teor Fiz* 35:479
43. Ohta T, Mimura M, Kobayashi R (1989) *Physica D* 34:115
44. Couillet P, Riera C, Tresser C (2004) *Chaos* 14:193
45. Reynolds WN, Pearson JE, Ponce-Dawson S (1994) *Phys Rev Lett* 72:2797
46. Or-Guil M, Bode M, Schenk CP, Purwins H-G (1998) *Phys Rev E* 57:6432
47. Elphick C, Meron E, Spiegel EA (1988) *Phys Rev Lett* 61:496
48. Malomed BA, Nepomnyashchy AA, Tribelsky MI (1990) *Phys Rev A* 42:7244
49. Ben-Jacob E, Brand HR, Dee G, Kramer L, Langer JS (1985) *Physica D* 14:348
50. Pomeau Y (1986) *Physica D* 23:3
51. Aranson IS, Malomed BA, Pismen LM, Tsimring LS (2000) *Phys Rev E* 62:R5
52. Hagberg A, Yochelis A, Yizhak H, Elphick C, Pismen LM, Meron E (2006) *Physica D* 217:186
53. Aranson IS, Kramer L (2002) *Rev Mod Phys* 74:99
54. van Saarloos W, Hohenberg PC (1992) *Physica D* 56:303
55. Brusch L, Torcini A, van Hecke M, Zimmermann MG, Bär M (2001) *Physica D* 160:127
56. Knobloch E, de Luca J (1990) *Nonlinearity* 3:975
57. Pismen LM (1986) *Dyn Stab Syst* 1:97
58. Hagan PS (1982) *SIAM J Appl Math* 42:762
59. Chaté H, Manneville P (1996) *Physica A* 224:348
60. Tyson JJ, Keener JP (1988) *Physica D* 32:327
61. Hakim V, Karma A (1999) *Phys Rev E* 60:5073
62. Barkley D (1994) *Phys Rev Lett* 72:164
63. Park J-S, Lee KJ (2002) *Phys Rev Lett* 88:224501
64. Keener JP (1988) *Physica D* 31:269
65. Henry H, Hakim V (2002) *Phys Rev E* 65:046235
66. Rousseau G, Chaté H, Kapral R (1998) *Phys Rev Lett* 80:5671
67. Alonso S, Kähler R, Mikhailov AS, Sagués F (2004) *Phys Rev E* 70:056201
68. Brito C, Aranson IS, Chaté H (2003) *Phys Rev Lett* 90:068301
69. Nepomnyashchy AA, Golovin AA, Gubareva V, Panfilov V (2004) *Physica D* 199:61
70. Pismen LM (1987) *Phys Rev Lett* 59:2740
71. Petrov V, Ouyang Q, Swinney HL (1997) *Nature* 388:655
72. Lin AL, Hagberg A, Meron E, Swinney HL (2004) *Phys Rev E* 69:066217
73. Couillet P, Emilsson K (1992) *Physica D* 61:119
74. Yochelis A, Elphick C, Hagberg A, Meron E (2004) *Physica D* 199:201
75. Gomila D, Colet P, San Miguel M, Oppo G-L (2007) *Eur Phys J Special Topics* 146:71
76. Gallego R, Walgraef D, San Miguel M, Toral R (2001) *Phys Rev E* 64:056218

## Books and Reviews

- Cross MC, Hohenberg P (1993) *Rev Mod Phys* 65:851
- Epstein IR, Pojman JA (1998) *Introduction to Nonlinear Chemical Dynamics*. Oxford University Press, New York
- Fife PC (1979) *Mathematical Aspects of Reacting and Diffusing Systems*. Springer, Berlin
- Haken H (2004) *Synergetics: Introduction and Advanced Topics*. Springer, Berlin
- Hoyle R (2006) *Pattern Formation*. Cambridge UP, Cambridge
- Kuramoto Y (1984) *Chemical Oscillations, Waves and Turbulence*. Springer, Berlin
- Manneville P (1990) *Dissipative Structures and Weak Turbulence*. Academic Press, San Diego
- Mikhailov AS (1991) *Foundations of Synergetics I: Distributed Active Systems II (with AY Loskutov): Complex Patterns*. Springer, Berlin
- Murray JD (1989) *Mathematical Biology*. Springer, Berlin; 2nd edn 1993; 3rd edn 2002/2003
- Pismen LM (1999) *Vortices in Nonlinear Fields*. Clarendon Press, Oxford
- Pismen LM (2006) *Patterns and Interfaces in Dissipative Dynamics*. Springer, Berlin
- Rabinovich MI, Ezersky AB, Weidman PD (2000) *The Dynamics of Patterns*. World Scientific, Singapore



- Walgraef D (1997) *Spatio-Temporal Pattern Formation*. Springer, New York
- Winfree AT (1987) *When Time Breaks Down*. Princeton University Press, Princeton

## Pedestrian, Crowd and Evacuation Dynamics

DIRK HELBING<sup>1,2</sup>, ANDERS JOHANSSON<sup>1</sup>

<sup>1</sup> ETH Zurich, Zurich, Switzerland

<sup>2</sup> Institute for Advanced Study, Collegium Budapest, Budapest, Hungary

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Pedestrian Dynamics  
 Crowd Dynamics  
 Evacuation Dynamics  
 Future Directions  
 Acknowledgments  
 Bibliography

### Glossary

- Collective intelligence** Emergent functional behavior of a large number of people that results from *interactions* of individuals rather than from individual reasoning or global optimization.
- Crowd** Agglomeration of many people in the same area at the same time. The density of the crowd is assumed to be high enough to cause continuous interactions with or reactions to other individuals.
- Crowd turbulence** Unanticipated and unintended irregular motion of individuals into different directions due to strong and rapidly changing forces in crowds of extreme density.
- Emergence** Spontaneous establishment of a qualitatively new behavior through non-linear interactions of many objects or subjects.
- Evolutionary optimization** Gradual optimization based on the effect of frequently repeated random mutations and selection processes based on some success function ("fitness").
- Faster-is-slower effect** This term reflects the observation that certain processes (in evacuation situations, production, traffic dynamics, or logistics) take more time if performed at high speed. In other words, waiting can

often help to coordinate the activities of several competing units and to speed up the average progress.

**Freezing-by-heating effect** Noise-induced blockage effect caused by the breakdown of direction-segregated walking patterns (typically two or more "lanes" characterized by a uniform direction of motion). "Noise" means frequent variations of the walking direction due to nervousness or impatience in the crowd, e.g. also frequent overtaking maneuvers in dense, slowly moving crowds.

**Panic** Breakdown of ordered, cooperative behavior of individuals due to anxious reactions to a certain event. Often, panic is characterized by attempted escape of many individuals from a real or perceived threat in situations of a perceived struggle for survival, which may end up in trampling or crushing of people in a crowd.

**Self-organization** Spontaneous organization (i. e. formation of ordered patterns) not induced by initial or boundary conditions, by regulations or constraints. Self-organization is a result of non-linear interactions between many objects or subjects, and it often causes different kinds of spatio-temporal patterns of motion.

**Social force** Vector describing acceleration or deceleration effects that are caused by social interactions rather than by physical interactions or fields.

### Definition of the Subject

The modeling of pedestrian motion is of great theoretical and practical interest. Recent experimental efforts have revealed quantitative details of pedestrian interactions, which have been successfully cast into mathematical equations. Furthermore, corresponding computer simulations of large numbers of pedestrians have been compared with the empirically observed dynamics of crowds. Such studies have led to a deeper understanding of how collective behavior on a macroscopic scale emerges from individual human interactions. Interestingly enough, the non-linear interactions of pedestrians lead to various complex, spatio-temporal pattern-formation phenomena. This includes the emergence of lanes of uniform walking direction, oscillations of the pedestrian flow at bottlenecks, and the formation of stripes in two intersecting flows. Such self-organized patterns of motion demonstrate that efficient, "intelligent" collective dynamics can be based on simple, local interactions. Under extreme conditions, however, coordination may break down, giving rise to critical crowd conditions. Examples are "freezing-by-heating" and "faster-is-slower" effects, but also the transition to "turbulent" crowd dynamics. These observations have im-



portant implications for the optimization of pedestrian facilities, in particular for evacuation situations.

## Introduction

The emergence of new, functional or complex collective behaviors in social systems has fascinated many scientists. One of the primary questions in this field is how cooperation or coordination patterns originate based on elementary individual interactions. While one could think that these are a result of intelligent human actions, it turns out that much simpler models assuming automatic responses can reproduce the observations very well. This suggests that humans are using their intelligence primarily for more complicated tasks, but also that simple interactions can lead to intelligent patterns of motion. Of course, it is reasonable to assume that these interactions are the result of a previous learning process that has optimized the automatic response in terms of minimizing collisions and delays. This, however, seems to be sufficient to explain most observations.

In this contribution, we will start with a short history of pedestrian modeling and, then, introduce a simplified model of pedestrian interactions, the “social force model”. Furthermore, we will discuss its calibration using video tracking data. Next, we will turn to the subject of crowd dynamics, as one typically finds the formation of large-scale spatio-temporal patterns of motion, when many pedestrians interact with each other. These patterns will be discussed in some detail before we will turn to evacuation situations and cases of extreme densities, where one can sometimes observe the breakdown of coordination. Finally, we will address possibilities to design improved pedestrian facilities, using special evolutionary algorithms.

## Pedestrian Dynamics

### Short History of Pedestrian Modeling

Pedestrians have been empirically studied for more than four decades [1,2,3]. The evaluation methods initially applied were based on direct observation, photographs, and time-lapse films. For a long time, the main goal of these studies was to develop a *level-of-service concept* [4], *design elements* of pedestrian facilities [5,6,7,8], or *planning guidelines* [9,10]. The latter have usually the form of *regression relations*, which are, however, not very well suited for the prediction of pedestrian flows in pedestrian zones and buildings with an exceptional architecture, or in challenging evacuation situations. Therefore, a number of simulation models have been proposed, e.g. *queuing mod-*

*els* [11], *transition matrix models* [12], and *stochastic models* [13], which are partly related to each other. In addition, there are models for the *route choice behavior* of pedestrians [14,15].

None of these concepts adequately takes into account the self-organization effects occurring in pedestrian crowds. These are the subject of recent experimental studies [8,16,17,18,19,20]. Most pedestrian models, however, were formulated before. A first modeling approach that appears to be suited to reproduce spatio-temporal patterns of motion was proposed by Henderson [21], who conjectured that pedestrian crowds behave similar to gases or fluids (see also [22]). This could be partially confirmed, but a realistic gas-kinetic or fluid-dynamic theory for pedestrians must contain corrections due to their particular interactions (i.e. avoidance and deceleration maneuvers) which, of course, do not obey momentum and energy conservation. Although such a theory can be actually formulated [23,24], for practical applications a direct simulation of *individual* pedestrian motion is favorable, since this is more flexible. As a consequence, pedestrian research mainly focuses on *agent-based models* of pedestrian crowds, which also allow one to consider local coordination problems. The “social force model” [25,26] is maybe the most well-known of these models, but we also like to mention *cellular automata* of pedestrian dynamics [27,28,29,30,31,32,33] and *AI-based models* [34,35].

### The Social Force Concept

In the following, we shall shortly introduce the social force concept, which reproduces most empirical observations in a simple and natural way. Human behavior often seems to be “chaotic”, irregular, and unpredictable. So, why and under what conditions can we model it by means of forces? First of all, we need to be confronted with a phenomenon of motion in some (quasi-)continuous space, which may be also an abstract behavioral space such as an opinion scale [36]. Moreover, it is favorable to have a system where the fluctuations due to unknown influences are not large compared to the systematic, deterministic part of motion. This is usually the case in pedestrian traffic, where people are confronted with standard situations and react “automatically” rather than taking complicated decisions, e.g. if they have to evade others.

This “automatic” behavior can be interpreted as the result of a *learning process* based on trial and error [37], which can be simulated with *evolutionary algorithms* [38]. For example, pedestrians have a preferred side of walking, since an asymmetrical avoidance behavior turns out to be profitable [25,37]. The related *formation of a behav-*

ioral convention can be described by means of *evolutionary game theory* [25,39].

Another requirement is the vectorial additivity of the separate force terms reflecting different environmental influences. This is probably an approximation, but there is some experimental evidence for it. Based on quantitative measurements for animals and test persons subject to separately or simultaneously applied stimuli of different nature and strength, one could show that the behavior in conflict situations can be described by a superposition of forces [40,41]. This fits well into a concept by Lewin [42], according to which behavioral changes are guided by so-called *social fields* or *social forces*, which has later on been put into mathematical terms [25,43]. In some cases, social forces, which determine the amount and direction of systematic behavioral changes, can be expressed as gradients of dynamically varying potentials, which reflect the social or behavioral fields resulting from the interactions of individuals. Such a social force concept was applied to opinion formation and migration [43], and it was particularly successful in the description of collective pedestrian behavior [8,25,26,37].

For reliable simulations of pedestrian crowds, we do not need to know whether a certain pedestrian, say, turns to the right at the next intersection. It is sufficient to have a good estimate what percentage of pedestrians turns to the right. This can be either empirically measured or estimated by means of route choice models [14]. In some sense, the uncertainty about the individual behaviors is averaged out at the macroscopic level of description. Nevertheless, we will use the more flexible microscopic simulation approach based on the social force concept. According to this, the temporal change of the location  $\mathbf{r}_\alpha(t)$  of pedestrian  $\alpha$  obeys the equation of motion

$$\frac{d\mathbf{r}_\alpha(t)}{dt} = \mathbf{v}_\alpha(t). \quad (1)$$

Moreover, if  $\mathbf{f}_\alpha(t)$  denotes the sum of social forces influencing pedestrian  $\alpha$  and if  $\xi_\alpha(t)$  are individual fluctuations reflecting unsystematic behavioral variations, the velocity changes are given by the *acceleration equation*

$$\frac{d\mathbf{v}_\alpha}{dt} = \mathbf{f}_\alpha(t) + \xi_\alpha(t). \quad (2)$$

A particular advantage of this approach is that we can take into account the flexible usage of space by pedestrians, requiring a continuous treatment of motion. It turns out that this point is essential to reproduce the empirical observations in a natural and robust way, i. e. without having to adjust the model to each single situation and measurement

site. Furthermore, it is interesting to note that, if the fluctuation term is neglected, the social force model can be interpreted as a particular *differential game*, i. e. its dynamics can be derived from the minimization of a special utility function [44].

### Specification of the Social Force Model

The social force model for pedestrians assumes that each individual  $\alpha$  is trying to move in a desired direction  $\mathbf{e}_\alpha^0$  with a desired speed  $v_\alpha^0$ , and that it adapts the actual velocity  $\mathbf{v}_\alpha$  to the desired one,  $\mathbf{v}_\alpha^0 = v_\alpha^0 \mathbf{e}_\alpha^0$ , within a certain relaxation time  $\tau_\alpha$ . The systematic part  $\mathbf{f}_\alpha(t)$  of the acceleration force of pedestrian  $\alpha$  is then given by

$$\mathbf{f}_\alpha(t) = \frac{1}{\tau_\alpha} (v_\alpha^0 \mathbf{e}_\alpha^0 - \mathbf{v}_\alpha) + \sum_{\beta (\neq \alpha)} \mathbf{f}_{\alpha\beta}(t) + \sum_i \mathbf{f}_{\alpha i}(t), \quad (3)$$

where the terms  $\mathbf{f}_{\alpha\beta}(t)$  and  $\mathbf{f}_{\alpha i}(t)$  denote the repulsive forces describing attempts to keep a certain safety distance to other pedestrians  $\beta$  and obstacles  $i$ . In very crowded situations, additional physical contact forces come into play (see Subsect. “[Force Model for Panicking Pedestrians](#)”). Further forces may be added to reflect attraction effects between members of a group or other influences. For details see [37].

First, we will assume a simplified interaction force of the form

$$\mathbf{f}_{\alpha\beta}(t) = \mathbf{f}(d_{\alpha\beta}(t)), \quad (4)$$

where  $d_{\alpha\beta} = \mathbf{r}_\alpha - \mathbf{r}_\beta$  is the distance vector pointing from pedestrian  $\beta$  to  $\alpha$ . Angular-dependent shielding effects may be furthermore taken into account by a prefactor describing the anisotropic reaction to situations in front of as compared to behind a pedestrian [26,45], see Subsect. “[Angular Dependence](#)”. However, we will start with a **circular specification** of the distance-dependent interaction force,

$$\mathbf{f}(d_{\alpha\beta}) = A_\alpha e^{-d_{\alpha\beta}/B_\alpha} \frac{d_{\alpha\beta}}{\|d_{\alpha\beta}\|}, \quad (5)$$

where  $d_{\alpha\beta} = \|d_{\alpha\beta}\|$  is the distance. The parameter  $A_\alpha$  reflects the *interaction strength*, and  $B_\alpha$  corresponds to the *interaction range*. While the dependence on  $\alpha$  explicitly allows for a dependence of these parameters on the single individual, we will assume a homogeneous population, i. e.  $A_\alpha = A$  and  $B_\alpha = B$  in the following. Otherwise, it would be hard to collect enough data for parameter calibration.

**Elliptical Specification** Note that it is possible to express Eq. (5) as gradient of an exponentially decaying potential

$V_{\alpha\beta}$ . This circumstance can be used to formulate a generalized, elliptical interaction force via the potential

$$V_{\alpha\beta}(b_{\alpha\beta}) = AB e^{-b_{\alpha\beta}/B}, \quad (6)$$

where the variable  $b_{\alpha\beta}$  denotes the semi-minor axis  $b_{\alpha\beta}$  of the elliptical equipotential lines. This has been specified according to

$$2b_{\alpha\beta} = \sqrt{\frac{(\|\mathbf{d}_{\alpha\beta}\| + \|\mathbf{d}_{\alpha\beta} - (\mathbf{v}_\beta - \mathbf{v}_\alpha)\Delta t\|)^2}{-\|(\mathbf{v}_\beta - \mathbf{v}_\alpha)\Delta t\|^2}}, \quad (7)$$

so that both pedestrians  $\alpha$  and  $\beta$  are treated symmetrically. The repulsive force is related to the above potential via

$$\begin{aligned} f_{\alpha\beta}(\mathbf{d}_{\alpha\beta}) &= -\nabla_{\mathbf{d}_{\alpha\beta}} V_{\alpha\beta}(b_{\alpha\beta}) \\ &= -\frac{dV_{\alpha\beta}(b_{\alpha\beta})}{db_{\alpha\beta}} \nabla_{\mathbf{d}_{\alpha\beta}} b_{\alpha\beta}(\mathbf{d}_{\alpha\beta}), \end{aligned} \quad (8)$$

where  $\nabla_{\mathbf{d}_{\alpha\beta}}$  represents the gradient with respect to  $\mathbf{d}_{\alpha\beta}$ . Considering the chain rule,  $\|\mathbf{z}\| = \sqrt{z^2}$ , and  $\nabla_{\mathbf{z}}\|\mathbf{z}\| = \mathbf{z}/\sqrt{z^2} = \mathbf{z}/\|\mathbf{z}\|$ , this leads to the explicit formula

$$\begin{aligned} f_{\alpha\beta}(\mathbf{d}_{\alpha\beta}) &= Ae^{-b_{\alpha\beta}/B} \cdot \frac{\|\mathbf{d}_{\alpha\beta}\| + \|\mathbf{d}_{\alpha\beta} - \mathbf{y}_{\alpha\beta}\|}{2b_{\alpha\beta}} \\ &\quad \cdot \frac{1}{2} \left( \frac{\mathbf{d}_{\alpha\beta}}{\|\mathbf{d}_{\alpha\beta}\|} + \frac{\mathbf{d}_{\alpha\beta} - \mathbf{y}_{\alpha\beta}}{\|\mathbf{d}_{\alpha\beta} - \mathbf{y}_{\alpha\beta}\|} \right) \end{aligned} \quad (9)$$

with  $\mathbf{y}_{\alpha\beta} = (\mathbf{v}_\beta - \mathbf{v}_\alpha)\Delta t$ . We used  $\Delta t = 0.5$  s. For  $\Delta t = 0$ , we regain the expression of Eq. (5).

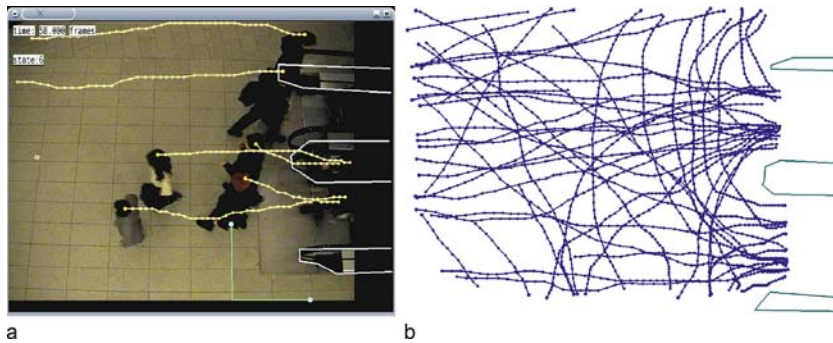
The elliptical specification has two major advantages compared to the circular one: First, the interactions depend not only on the distance, but also on the relative velocity. Second, the repulsive force is not strictly directed from pedestrian  $\beta$  to pedestrian  $\alpha$ , but has a lateral component. As a consequence, this leads to less confrontative,

smoother (“sliding”) evading maneuvers. Note that further velocity-dependent specifications of pedestrian interaction forces have been proposed [7,26], but we will restrict to the above specifications, as these are sufficient to demonstrate the method of evolutionary model calibration.

### Evolutionary Calibration with Video Tracking Data

For parameter calibration, several video recordings of pedestrian crowds in different natural environments have been used. The dimensions of the recorded areas were known, and the floor tiling or environment provided something like a “coordinate system”. The heads were automatically determined by searching for round moving structures, and the accuracy of tracking was improved by comparing actual with linearly extrapolated positions (so it would not happen so easily that the algorithm interchanged or “lost” close by pedestrians). The trajectories of the heads were then projected on two-dimensional space in a way correcting for distortion by the camera perspective. A representative plot of the resulting trajectories is shown in Fig. 1. Note that trajectory data have been obtained with infra-red sensors [47] or video cameras [48,49] for several years now, but algorithms that can simultaneously handle more than one thousand pedestrians have become available only recently [87].

For model calibration, it is recommended to use a hybrid method fusing empirical trajectory data and microscopic simulation data of pedestrian movement in space. In corresponding algorithms, a virtual pedestrian is assigned to each tracked pedestrian in the simulation domain. One then starts a simulation for a time period  $T$  (e.g. 1.5 s), in which one pedestrian  $\alpha$  is moved according to a simulation of the social force model, while the others are moved exactly according to the trajectories ex-



**Pedestrian, Crowd and Evacuation Dynamics, Figure 1**

Video tracking used to extract the trajectories of pedestrians from video recordings close to two escalators (after [45]). **a** Illustration of the tracking of pedestrian heads. **b** Resulting trajectories after being transformed onto the two-dimensional plane

tracted from the videos. This procedure is performed for all pedestrians  $\alpha$  and for several different starting times  $t$ , using a fixed parameter set for the social force model.

Each simulation run is performed according to the following scheme:

1. Define a starting point and calculate the state (position  $\mathbf{r}_\alpha$ , velocity  $\mathbf{v}_\alpha$ , and acceleration  $\mathbf{a}_\alpha = d\mathbf{v}_\alpha/dt$ ) for each pedestrian  $\alpha$ .
2. Assign a desired speed  $v_\alpha^0$  to each pedestrian, e.g. the maximum speed during the pedestrian tracking time. This is sufficiently accurate, if the overall pedestrian density is not too high and the desired speed is constant in time.
3. Assign a desired goal point for each pedestrian, e.g. the end point of the trajectory.
4. Given the tracked motion of the surrounding pedestrians  $\beta$ , simulate the trajectory of pedestrian  $\alpha$  over a time period  $T$  based on the social force model, starting at the actual location  $\mathbf{r}_\alpha(t)$ .

After each simulation run, one determines the relative distance error

$$\frac{\|\mathbf{r}_\alpha^{\text{simulated}}(t+T) - \mathbf{r}_\alpha^{\text{tracked}}(t+T)\|}{\|\mathbf{r}_\alpha^{\text{tracked}}(t+T) - \mathbf{r}_\alpha^{\text{tracked}}(t)\|}. \quad (10)$$

After averaging the relative distance errors over the pedestrians  $\alpha$  and starting times  $t$ , 1 minus the result can be taken as measure of the goodness of fit (the “fitness”) of the parameter set used in the pedestrian simulation. Hence, the best possible value of the “fitness” is 1, but any deviation from the real pedestrian trajectories implies lower values.

One result of such a parameter optimization is that, for each video, there is a broad range of parameter combinations of  $A$  and  $B$  which perform almost equally well [45]. This allows one to apply additional goal functions in the parameter optimization, e.g. to determine among the best performing parameter values such parameter combinations, which perform well for *several* video recordings, using a fitness function which equally weights the fitness reached in each single video. This is how the parameter values listed in Table 1 were determined. It turns out that, in order to reach a good model performance, the pedestrian interaction force must be specified velocity dependent, as in the elliptical model.

Note that our evolutionary fitting method can be also used to determine interaction laws without prespecified interaction functions. For example, one can obtain the distance dependence of pedestrian interactions without a prespecified function. For this, one adjusts the values of the

**Pedestrian, Crowd and Evacuation Dynamics, Table 1**

Interaction strength  $A$  and interaction range  $B$  resulting from our evolutionary parameter calibration for the circular and elliptical specification of the interaction forces between pedestrians (see main text). The calibration was based on three different video recordings, one for low crowd density, one for medium, and one for high density. The parameter values are specified as mean value  $\pm$  standard deviation. The best fitness value obtained with the elliptical specification for the video with the lowest crowd density was as high as 0.9

Model	A	B	“Fitness”
Extrapolation	0	–	0.34
Circular	$0.11 \pm 0.06$	$0.84 \pm 0.63$	0.35
Elliptical	$4.30 \pm 3.91$	$1.07 \pm 1.35$	0.53

force at given distances  $d_k = kd_1$  (with  $k \in \{1, 2, 3, \dots\}$ ) in an evolutionary way. To get some smoothness, linear interpolation is applied. The resulting fit curve is presented in Fig. 2 (left). It turns out that the empirical dependence of the force with distance can be well fitted by an exponential decay.

### Angular Dependence

A closer study of pedestrian interactions reveals that these are not isotropic, but dependent on the angle  $\varphi_{\alpha\beta}$  of the encounter, which is given by the formula

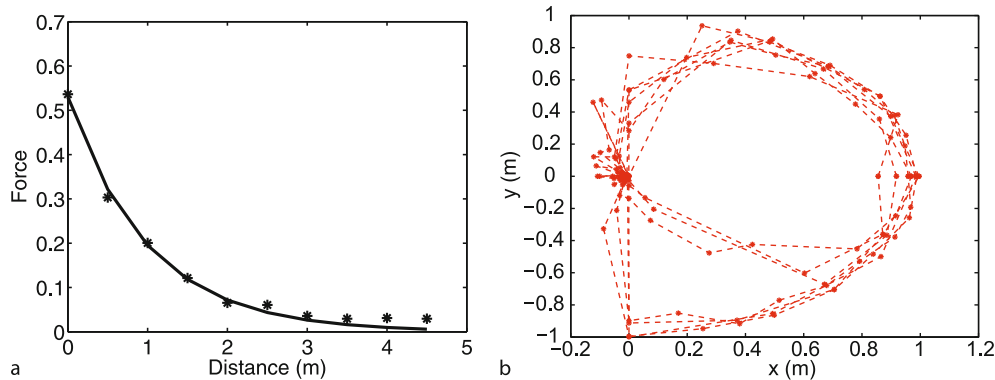
$$\cos(\varphi_{\alpha\beta}) = \frac{\mathbf{v}_\alpha}{\|\mathbf{v}_\alpha\|} \cdot \frac{-\mathbf{d}_{\alpha\beta}}{\|\mathbf{d}_{\alpha\beta}\|}. \quad (11)$$

Generally, pedestrians show little response to pedestrians behind them. This can be reflected by an angular-dependent prefactor  $w(\varphi_{\alpha\beta})$  of the interaction force [45]. Empirical results are represented in Fig. 2 (right). Reasonable results are obtained for the following specification of the prefactor:

$$w(\varphi_{\alpha\beta}(t)) = \left( \lambda_\alpha + (1 - \lambda_\alpha) \frac{1 + \cos(\varphi_{\alpha\beta})}{2} \right), \quad (12)$$

where  $\lambda_\alpha$  with  $0 \leq \lambda_\alpha \leq 1$  is a parameter which grows with the strength of interactions from behind. An evolutionary parameter optimization gives values  $\lambda \approx 0.1$  [45], i. e. a strong anisotropy. With such an angle-dependent prefactor, the “fitness” of the elliptical force increases from 0.53 to 0.61, when calibrated to the same set of videos. Other angular-dependent specifications split up the interaction force between pedestrians into a component against the direction of motion and another one perpendicular to it. Such a description allows for even smoother avoidance maneuvers.





**Pedestrian, Crowd and Evacuation Dynamics, Figure 2**

Results of an evolutionary fitting of pedestrian interactions. **a** Empirically determined distance dependence of the interaction force between pedestrians (after [45]). An exponential decay fits the empirical data quite well. The dashed fit curve corresponds to Eq. (5) with the parameters  $A = 0.53$  and  $B = 1.0$ . **b** Angular dependence of the influence of other pedestrians. The direction along the positive  $x$ -axis corresponds to the walking direction of pedestrians,  $y$  to the perpendicular direction

## Crowd Dynamics

### Analogies with Gases, Fluids, and Granular Media

When the density is low, pedestrians can move freely, and the observed crowd dynamics can be partially compared with the behavior of gases. At medium and high densities, however, the motion of pedestrian crowds shows some striking analogies with the motion of fluids:

1. Footprints of pedestrians in snow look similar to streamlines of fluids [15].
2. At borderlines between opposite directions of walking one can observe “viscous fingering” [50,51].
3. The emergence of pedestrian streams through standing crowds [7,37,52] appears analogous to the formation of river beds [53,54].

At high densities, however, the observations have rather analogies with driven granular flows. This will be elaborated in more detail in Sects. “Force Model for Panicking Pedestrians” and “Collective Phenomena in Panic Situations”. In summary, one could say that fluid-dynamic analogies work reasonably well in normal situations, while granular aspects dominate at extreme densities. Nevertheless, the analogy is limited, since the self-driven motion and the violation of momentum conservation imply special properties of pedestrian flows. For example, one usually does not observe eddies, which typically occur in regular fluids at high enough Reynolds numbers.

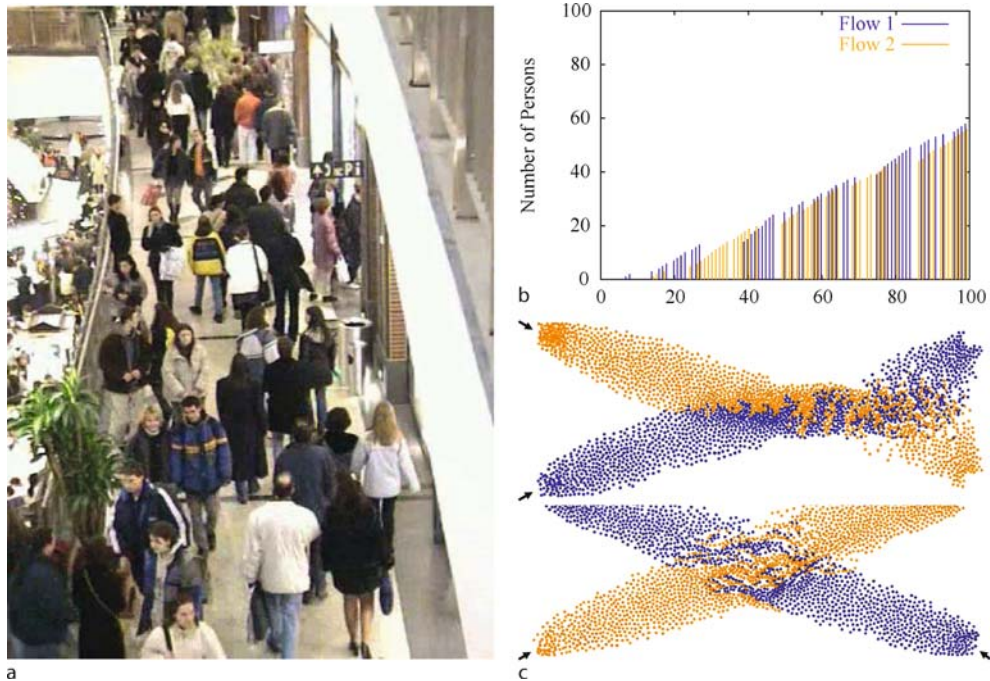
### Self-Organization of Pedestrian Crowds

Despite its simplifications, the social force model of pedestrian dynamics describes a lot of observed phenomena

quite realistically. Especially, it allows one to explain various self-organized spatio-temporal patterns that are not externally planned, prescribed, or organized, e.g. by traffic signs, laws, or behavioral conventions. Instead, the spatio-temporal patterns discussed below emerge due to the non-linear interactions of pedestrians even without assuming strategic considerations, communication, or imitative behavior of pedestrians. Despite this, we may still interpret the forming cooperation patterns as phenomena that establish social order on short time scales. It is actually surprising that strangers coordinate with each other within seconds, if they have grown up in a similar environment. People from different countries, however, are sometimes irritated about local walking habits, which indicates that learning effects and cultural backgrounds still play a role in social interactions as simple as random pedestrian encounters. Rather than on particular features, however, in the following we will focus on the common, internationally reproducible observations.

**Lane Formation** In pedestrian flows one can often observe that oppositely moving pedestrians are forming lanes of uniform walking direction (see Fig. 3). This phenomenon even occurs when there is not a large distance to separate each other, e.g. on zebra crossings. However, the width of lanes increases (and their number decreases), if the interaction continues over longer distances (and if perturbations, e.g. by flows entering or leaving on the sides, are low; otherwise the phenomenon of lane formation may break down [55]).

Lane formation may be viewed as *segregation phenomenon* [56,57]. Although there is a weak preference for one side (with the corresponding behavioral convention



**Pedestrian, Crowd and Evacuation Dynamics, Figure 3**

Self-organization of pedestrian crowds. **a** Photograph of lanes formed in a shopping center. Computer simulations reproduce the self-organization of such lanes very well. **b** Evaluation of the cumulative number of pedestrians passing a bottleneck from different sides. One can clearly see that the narrowing is often passed by groups of people in an oscillatory way rather than one by one. **c** Multi-agent simulation of two crossing pedestrian streams, showing the phenomenon of stripe formation. This self-organized pattern allows pedestrians to pass the other stream without having to stop, namely by moving sideways in a forwardly moving stripe. (After [8])

depending on the country), the observations can only be well reproduced when repulsive pedestrian interactions are taken into account. The most relevant factor for the lane formation phenomenon is the higher relative velocity of pedestrians walking in opposite directions. Compared to people following each other, oppositely moving pedestrians have more frequent interactions until they have segregated into separate lanes by stepping aside whenever another pedestrian is encountered. The most long-lived patterns of motion are the ones which change the least. It is obvious that such patterns correspond to lanes, as they minimize the frequency and strength of avoidance maneuvers. Interestingly enough, as computer simulations show, lane formation occurs also when there is no preference for any side.

Lanes minimize frictional effects, accelerations, energy consumption, and delays in oppositely moving crowds. Therefore, one could say that they are a pattern reflecting “collective intelligence”. In fact, it is not possible for a single pedestrian to reach such a collective pattern of motion. Lane formation is a self-organized collaborative

pattern of motion originating from simple pedestrian interactions. Particularly in cases of no side preference, the system behavior cannot be understood by adding up the behavior of the single individuals. This is a typical feature of complex, self-organizing systems and, in fact, a widespread characteristics of social systems. It is worth noting, however, that it does not require a conscious behavior to reach forms of social organization like the segregation of oppositely moving pedestrians into lanes. This organization occurs automatically, although most people are not even aware of the existence of this phenomenon.

**Oscillatory Flows at Bottlenecks** At bottlenecks, bidirectional flows of moderate density are often characterized by oscillatory changes in the flow direction (see Fig. 3). For example, one can sometimes observe this at entrances of museums during crowded art exhibitions or at entrances of staff canteens during lunch time. While these oscillatory flows may be interpreted as an effect of friendly behavior (“you go first, please”), computer simulations of the social force model indicate that the collective behavior may again

be understood by simple pedestrian interactions. That is, oscillatory flows occur even in the absence of communication. Therefore, they may be viewed as another self-organization phenomenon, which again reduces frictional effects and delays. That is, oscillatory flows have features of “collective intelligence”.

While this may be interpreted as result of a learning effect in a large number of similar situations (a “repeated game”), our simulations suggest an even simpler, “many-particle” interpretation: Once a pedestrian is able to pass the narrowing, pedestrians with the same walking direction can easily follow. Hence, the number and “pressure” of waiting, “pushy” pedestrians on one side of the bottleneck becomes less than on the other side. This eventually increases their chance to occupy the passage. Finally, the “pressure difference” is large enough to stop the flow and turn the passing direction at the bottleneck. This reverses the situation, and eventually the flow direction changes again, giving rise to oscillatory flows.

**Stripe Formation in Intersecting Flows** In intersection areas, the flow of people often appears to be irregular or “chaotic”. In fact, it can be shown that there are several possible collective patterns of motion, among them rotary and oscillating flows. However, these patterns continuously compete with each other, and a temporarily dominating pattern is destroyed by another one after a short time. Obviously, there has not evolved any social convention that would establish and stabilize an ordered and efficient flow at intersections.

Self-organized patterns of motion, however, are found in situations where pedestrian flows cross each other only in two directions. In such situations, the phenomenon of stripe formation is observed [58]. Stripe formation allows two flows to penetrate each other without requiring the pedestrians to stop. For an illustration see Fig. 3. Like lanes, stripes are a segregation phenomenon, but not a stationary one. Instead, the stripes are density waves moving into the direction of the sum of the directional vectors of both intersecting flows. Naturally, the stripes extend sideways into the direction which is perpendicular to their direction of motion. Therefore, the pedestrians move forward *with* the stripes and sideways *within* the stripes. Lane formation corresponds to the particular case of stripe formation where both directions are exactly opposite. In this case, no intersection takes place, and the stripes do not move systematically. As in lane formation, stripe formation allows to minimize obstructing interactions and to maximize the average pedestrian speeds, i. e. simple, repulsive pedestrian interactions again lead to an “intelligent” collective behavior.

## Evacuation Dynamics

While the previous section has focused on the dynamics of pedestrian crowds in normal situations, we will now turn to the description of situations in which extreme crowd densities occur. Such situations may arise at mass events, particularly in cases of urgent egress. While most evacuations run relatively smoothly and orderly, the situation may also get out of control and end up in terrible crowd disasters (see Table 2). In such situations, one often speaks of “panic”, although, from a scientific standpoint, the use of this term is rather controversial. Here, however, we will not be interested in the question whether “panic” actually occurs or not. We will rather focus on the issue of crowd dynamics at high densities and under psychological stress.

## Evacuation and Panic Research

Computer models have been also developed for emergency and evacuation situations [32,60,61,62,63,64,65,66,67,68]. Most research into panic, however, has been of empirical nature (see, e. g. [69,70,71,72]), carried out by social psychologists and others.

With some exceptions, panic is observed in cases of scarce or dwindling resources [69,73], which are either required for survival or anxiously desired. They are usually distinguished into escape panic (“stampedes”, bank or stock market panic) and acquisitive panic (“crazes”, speculative manias) [74,75], but in some cases this classification is questionable [76].

It is often stated that panicking people are obsessed by short-term personal interests uncontrolled by social and cultural constraints [69,74]. This is possibly a result of the reduced attention in situations of fear [69], which also causes that options like side exits are mostly ignored [70]. It is, however, mostly attributed to social contagion [69,71,73,74,75,76,77,78,79,80,81], i. e., a transition from individual to mass psychology, in which individuals transfer control over their actions to others [75], leading to conformity [82]. This “herding behavior” is in some sense irrational, as it often leads to bad overall results like dangerous overcrowding and slower escape [70,75,76]. In this way, herding behavior can increase the fatalities or, more generally, the damage in the crisis faced.

The various socio-psychological theories for this contagion assume hypnotic effects, rapport, mutual excitation of a primordial instinct, circular reactions, social facilitation (see the summary by Brown [80]), or the emergence of normative support for selfish behavior [81]. Brown [80] and Coleman [75] add another explanation related to the prisoner’s dilemma [83,84] or common goods dilemma [85], showing that it is reasonable to make one’s

**Pedestrian, Crowd and Evacuation Dynamics, Table 2**

Incomplete list of major crowd disasters since 1970 after J. F. Dickie in [59], <http://www.crowddynamics.com/Main/Crowddisasters.html>, [http://SportsIllustrated.CNN.com/soccer/world/news/2000/07/09/stadium\\_disasters\\_ap/](http://SportsIllustrated.CNN.com/soccer/world/news/2000/07/09/stadium_disasters_ap/), and other internet sources, excluding fires, bomb attacks, and train or plane accidents. The number of injured people was usually a multiple of the fatalities

Date	Place	Venue	Deaths	Reason
1971	Ibroy, UK	Stadium	66	Collapse of barriers
1974	Cairo, Egypt	Stadium	48	Crowds break barriers
1982	Moscow, USSR	Stadium	340	Re-entering fans after last minute goal
1988	Katmandu, Nepal	Stadium	93	Stampede due to hailstorm
1989	Hillsborough, Sheffield, UK	Stadium	96	Fans trying to force their way into the stadium
1990	New York City	Bronx	87	Illegal happy land social club
1990	Mena, Saudi Arabia	Pedestrian Tunnel	1426	Overcrowding
1994	Mena, Saudi Arabia	Jamarat Bridge	266	Overcrowding
1996	Guatemala City, Guatemala	Stadium	83	Fans trying to force their way into the stadium
1998	Mena, Saudi Arabia		118	Overcrowding
1999	Kerala, India	Hindu Shrine	51	Collapse of parts of the shrine
1999	Minsk, Belarus	Subway Station	53	Heavy rain at rock concert
2001	Ghana, West Africa	Stadium	> 100	Panic triggered by tear gas
2004	Mena, Saudi Arabia	Jamarat Bridge	251	Overcrowding
2005	Wai, India	Religious Procession	150	Overcrowding (and fire)
2005	Bagdad, Iraque	Religious Procession	> 640	Rumors regarding suicide bomber
2005	Chennai, India	Disaster Area	42	Rush for flood relief supplies
2006	Mena, Saudi Arabia	Jamarat Bridge	363	Overcrowding
2006	Pilippines	Stadium	79	Rush for game show tickets
2006	Ibb, Yemen	Stadium	51	Rally for Yemeni president

subsequent actions contingent upon those of others. However, the socially favorable behavior of walking orderly is unstable, which normally gives rise to rushing by everyone. These thoughtful considerations are well compatible with many aspects discussed above and with the classical experiments by Mintz [73], which showed that jamming in escape situations depends on the reward structure (“payoff matrix”).

Nevertheless and despite of the frequent reports in the media and many published investigations of crowd disasters (see Table 2), a quantitative understanding of the observed phenomena in panic stampedes was lacking for a long time. In the following, we will close this gap.

### Situations of “Panic”

Panic stampede is one of the most tragic collective behaviors [71,72,73,74,75,77,78,79,80,81], as it often leads to the death of people who are either crushed or trampled down by others. While this behavior may be comprehensible in life-threatening situations like fires in crowded buildings [69,70], it is hard to understand in cases of a rush for good seats at a pop concert [76] or without any obvious reasons. Unfortunately, the frequency of such disasters is increasing (see Table 2), as growing population den-

sities combined with easier transportation lead to greater mass events like pop concerts, sport events, and demonstrations. Nevertheless, systematic empirical studies of panic [73,86] are rare [69,74,76], and there is a scarcity of quantitative theories capable of predicting crowd dynamics at extreme densities [32,60,61,64,65,68]. The following features appear to be typical [46,55]:

1. In situations of escape panic, individuals are getting nervous, i. e. they tend to develop blind actionism.
2. People try to move considerably faster than normal [9].
3. Individuals start pushing, and interactions among people become physical in nature.
4. Moving and, in particular, passing of a bottleneck frequently becomes incoordinated [73].
5. At exits, jams are building up [73]. Sometimes, intermittent flows or arching and clogging are observed [9], see Fig. 4.
6. The physical interactions in jammed crowds add up and can cause dangerous pressures up to 4,500 Newtons per meter [59,70], which can bend steel barriers or tear down brick walls.
7. The strength and direction of the forces acting in large crowds can suddenly change [87], pushing peo-





**Pedestrian, Crowd and Evacuation Dynamics, Figure 4**  
Panicking football fans trying to escape the football stadium in Sheffield. Because of a clogging effect, it is difficult to pass the open door

ple around in an uncontrollable way. This may cause people to fall.

8. Escape is slowed down by fallen or injured people turning into “obstacles”.
9. People tend to show herding behavior, i. e., to do what other people do [69,78].
10. Alternative exits are often overlooked or not efficiently used in escape situations [69,70].

The following quotations give a more personal impression of the conditions during crowd panic:

1. “They just kept pushin’ forward and they would just walk right on top of you, just trample over ya like you were a piece of the ground.” (After the panic at “The Who Concert Stampede” in Cincinnati.)
2. “People were climbin’ over people ta get in ... an’ at one point I almost started hittin’ ’em, because I could not believe the animal, animalistic ways of the people, you know, nobody cared.” (After the panic at “The Who Concert Stampede”.)
3. “Smaller people began passing out. I attempted to lift one girl up and above to be passed back ... After several tries I was unsuccessful and near exhaustion.” (After the panic at “The Who Concert Stampede”.)
4. “I couldn’t see the floor because of the thickness of the smoke.” (After the “Hilton Hotel Fire” in Las Vegas.)
5. “The club had two exits, but the young people had access to only one”, said Narend Singh, provincial minister for agriculture and environmental affairs. However, the club’s owner, Rajan Naidoo, said the club had four exits, and that all were open. “I think the children panicked and headed for the main entrance where they

initially came in,’ he said.” (After the “Durban Disco Stampede”.)

6. “At occupancies of about 7 persons per square meter the crowd becomes almost a fluid mass. Shock waves can be propagated through the mass, sufficient to ... propel them distances of 3 meters or more. ... People may be literally lifted out of their shoes, and have clothing torn off. Intense crowd pressures, exacerbated by anxiety, make it difficult to breathe, which may finally cause compressive asphyxia. The heat and the thermal insulation of surrounding bodies cause some to be weakened and faint. Access to those who fall is impossible. Removal of those in distress can only be accomplished by lifting them up and passing them overhead to the exterior of the crowd.” (J. Fruin in [88].)
7. “It was like a huge wave of sea gushing down on the pilgrims” (P. K. Abdul Ghafour, Arab News, after the sad crowd disaster in Mena on January 12, 2006).

### Force Model for Panicking Pedestrians

Additional, physical interaction forces  $f_{\alpha\beta}^{\text{ph}}$  come into play when pedestrians get so close to each other that they have physical contact (i. e.  $d_{\alpha\beta} < r_{\alpha\beta} = r_{\alpha} + r_{\beta}$ , where  $r_{\alpha}$  means the “radius” of pedestrian  $\alpha$ ). In this case, which is mainly relevant to panic situations, we assume also a “body force”  $k(r_{\alpha\beta} - d_{\alpha\beta})\mathbf{n}_{\alpha\beta}$  counteracting body compression and a “sliding friction force”  $\kappa(r_{\alpha\beta} - d_{\alpha\beta})\Delta\mathbf{v}_{\beta\alpha}^t \mathbf{t}_{\alpha\beta}$  impeding relative tangential motion. Inspired by the formulas for granular interactions [89,90], we assume

$$\mathbf{f}_{\alpha\beta}^{\text{ph}}(t) = k\Theta(r_{\alpha\beta} - d_{\alpha\beta})\mathbf{n}_{\alpha\beta} + \kappa\Theta(r_{\alpha\beta} - d_{\alpha\beta})\Delta\mathbf{v}_{\beta\alpha}^t \mathbf{t}_{\alpha\beta}, \quad (13)$$

where the function  $\Theta(z)$  is equal to its argument  $z$ , if  $z \geq 0$ , otherwise 0. Moreover,  $\mathbf{t}_{\alpha\beta} = (-n_{\alpha\beta}^2, n_{\alpha\beta}^1)$  means the tangential direction and  $\Delta\mathbf{v}_{\beta\alpha}^t = (\mathbf{v}_{\beta} - \mathbf{v}_{\alpha}) \cdot \mathbf{t}_{\alpha\beta}$  the tangential velocity difference, while  $k$  and  $\kappa$  represent large constants. (Strictly speaking, friction effects already set in before pedestrians touch each other, because of the psychological tendency not to pass other individuals with a high relative velocity, when the distance is small.)

The interactions with the boundaries of walls and other obstacles are treated analogously to pedestrian interactions, i. e., if  $d_{\alpha i}(t)$  means the distance to obstacle or boundary  $i$ ,  $\mathbf{n}_{\alpha i}(t)$  denotes the direction perpendicular to it, and  $\mathbf{t}_{\alpha i}(t)$  the direction tangential to it, the corresponding interaction force with the boundary reads

$$\mathbf{f}_{\alpha i} = \{A_{\alpha} \exp[(r_{\alpha} - d_{\alpha i})/B_{\alpha}] + k\Theta(r_{\alpha} - d_{\alpha i})\} \times \mathbf{n}_{\alpha i} - \kappa\Theta(r_{\alpha} - d_{\alpha i})(\mathbf{v}_{\alpha} \cdot \mathbf{t}_{\alpha i})\mathbf{t}_{\alpha i}. \quad (14)$$

Finally, fire fronts are reflected by repulsive social forces similar those describing walls, but they are much stronger. The physical interactions, however, are qualitatively different, as people reached by the fire front become injured and immobile ( $v_\alpha = 0$ ).

### Collective Phenomena in Panic Situations

In panic situations (e. g. in some cases of emergency evacuation) the following characteristic features of pedestrian behavior are often observed:

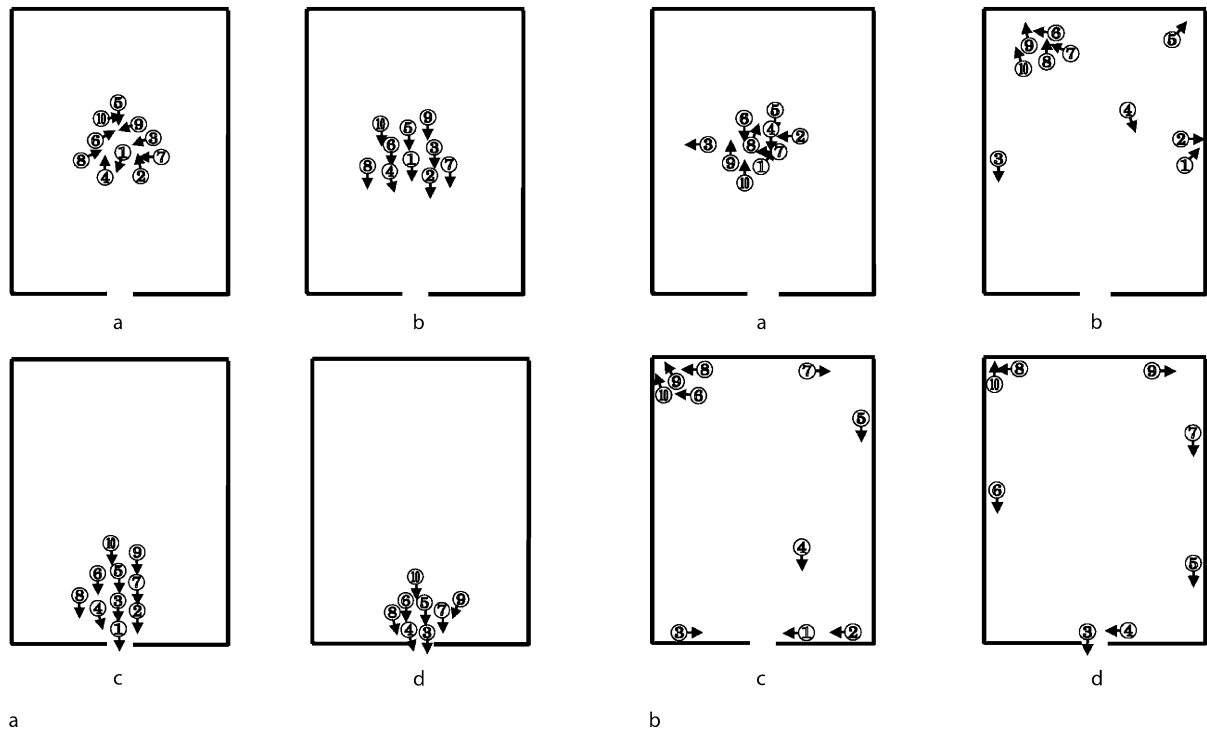
1. People are getting nervous, resulting in a higher level of fluctuations.
2. They are trying to escape from the source of panic, which can be reflected by a significantly higher desired velocity  $v_\alpha^0$ .
3. Individuals in complex situations, who do not know what is the right thing to do, orient at the actions of their neighbors, i. e. they tend to do what other people do. We will describe this by an additional herding interaction.

We will now discuss the fundamental collective effects which fluctuations, increased desired velocities, and herd-

ing behavior can have. In contrast to other approaches, we do not assume or imply that individuals in panic or emergency situations would behave relentless and asocial, although they sometimes do.

**Herding and Ignorance of Available Exits** If people are not sure what is the best thing to do, there is a tendency to show a “herding behavior”, i. e. to imitate the behavior of others. Fashion, hypes and trends are examples for this. The phenomenon is also known from stock markets, and particularly pronounced when people are anxious. Such a situation is, for example, given if people need to escape from a smoky room. There, the evacuation dynamics is very different from normal leaving (see Fig. 5).

Under normal visibility, everybody easily finds an exit and uses more or less the shortest path. However, when the exit cannot be seen, evacuation is much less efficient and may take a long time. Most people tend to walk relatively straight into the direction in which they suspect an exit, but in most cases, they end up at a wall. Then, they usually move along it in one of the two possible directions, until they finally find an exit [18]. If they encounter others, there is a tendency to take a decision for one direction



**Pedestrian, Crowd and Evacuation Dynamics, Figure 5**

**a** Normal leaving of a room, when the exit is well visible. **b** Escape from a room with no visibility, e. g. due to dense smoke or a power blackout. (After [18])

and move collectively. Also in case of acoustic signals, people may be attracted into the same direction. This can lead to over-crowded exits, while other exits are ignored. The same can happen even for normal visibility, when people are not well familiar with their environment and are not aware of the directions of the emergency exits.

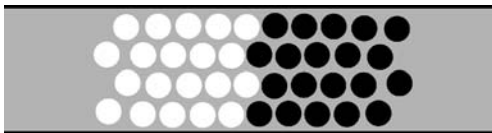
Computer simulations suggest that neither individualistic nor herding behavior performs well [46]. Pure individualistic behavior means that each pedestrian finds an exit only accidentally, while pure herding behavior implies that the complete crowd is eventually moving into the same and probably congested direction, so that available emergency exits are not efficiently used. Optimal chances of survival are expected for a certain mixture of individualistic and herding behavior, where individualism allows *some* people to detect the exits and herding guarantees that successful solutions are imitated by small groups of others [46].

**“Freezing by Heating”** Another effect of getting nervous has been investigated in [55]. Let us assume the individual fluctuation strength is given by

$$\eta_\alpha = (1 - n_\alpha)\eta_0 + n_\alpha\eta_{\max}, \quad (15)$$

where  $n_\alpha$  with  $0 \leq n_\alpha \leq 1$  measures the nervousness of pedestrian  $\alpha$ . The parameter  $\eta_0$  means the normal and  $\eta_{\max}$  the maximum fluctuation strength. It turns out that, at sufficiently high pedestrian densities, lanes are destroyed by increasing the fluctuation strength (which is analogous to the temperature). However, instead of the expected transition from the “fluid” lane state to a disordered, “gaseous” state, a solid state is formed. It is characterized by an at least temporarily blocked, “frozen” situation so that one calls this paradoxical transition “*freezing by heating*” (see Fig. 6). Notably enough, the blocked state has a *higher* degree of order, although the internal energy is *increased* [55].

The preconditions for this unusual freezing-by-heating transition are the driving term  $v_\alpha^0 e_\alpha^0 / \tau_\alpha$  and the dissipative friction  $-v_\alpha / \tau_\alpha$ , while the sliding friction force is not required. Inhomogeneities in the channel diameter



**Pedestrian, Crowd and Evacuation Dynamics, Figure 6**  
Result of the noise-induced formation of a “frozen” state in a (periodic) corridor used by oppositely moving pedestrians (after [55])

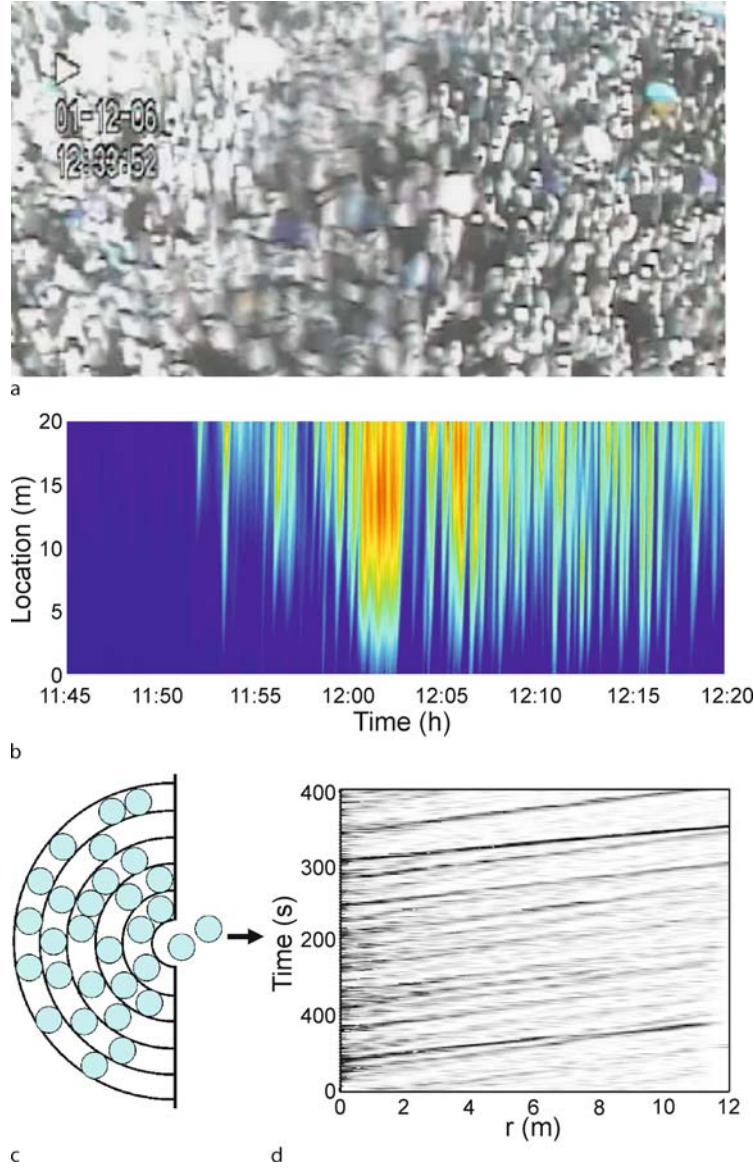
or other impurities which temporarily slow down pedestrians can further this transition at the respective places. Finally note that a transition from fluid to blocked pedestrian counter flows is also observed, when a critical density is exceeded [31,55].

**Intermittent Flows, Faster-Is-Slower Effect, and “Phantom Panic”** If the overall flow towards a bottleneck is higher than the overall outflow from it, a pedestrian queue emerges [91]. In other words, a waiting crowd is formed upstream of the bottleneck. High densities can result, if people keep heading forward, as this eventually leads to higher and higher compressions. Particularly critical situations may occur if the arrival flow is much higher than the departure flow, especially if people are trying to get towards a strongly desired goal (“acquisitive panic”) or away from a perceived source of danger (“escape panic”) with an increased driving force  $v_\alpha^0 e_\alpha^0 / \tau$ . In such situations, the high density causes coordination problems, as several people compete for the same few gaps. This typically causes body interactions and frictional effects, which can slow down crowd motion or evacuation (“*faster is slower effect*”).

A possible consequence of these coordination problems are intermittent flows. In such cases, the outflow from the bottleneck is not constant, but it is typically interrupted. While one possible origin of the intermittent flows are clogging and arching effects as known from granular flows through funnels or hoppers [89,90], stop-and-go waves have also been observed in more than 10 meter wide streets and in the 44 meters wide entrance area to the Jamarat Bridge during the pilgrimage in January 12, 2006 [87], see Fig. 7. Therefore, it seems to be important that people do not move continuously, but have minimum strides [25]. That is, once a person is stopped, he or she will not move until some space opens up in front. However, increasing impatience will eventually reduce the minimum stride, so that people eventually start moving again, even if the outflow through the bottleneck is stopped. This will lead to a further compression of the crowd.

In the worst case, such behavior can trigger a “*phantom panic*”, i. e. a crowd disaster *without* any serious reasons (e. g., in Moscow, 1982). For example, due to the “faster-is-slower effect” panic can be triggered by small pedestrian counterflows [70], which cause delays to the crowd intending to leave. Consequently, stopped pedestrians in the back, who do not see the reason for the temporary slowdown, are getting impatient and pushy. In accordance with observations [7,25], one may model this by increasing the desired velocity, for example, by the formula

$$v_\alpha^0(t) = [1 - n_\alpha(t)]v_\alpha^0(0) + n_\alpha(t)v_\alpha^{\max}. \quad (16)$$



**Pedestrian, Crowd and Evacuation Dynamics, Figure 7**

**a** Long-term photograph showing stop-and-go waves in a densely packed street. While stopped people appear relatively sharp, people moving from right to left have a fuzzy appearance. Note that gaps propagate from *right to left*. **b** Empirically observed stop-and-go waves in front of the entrance to the Jamarat Bridge on January 12, 2006 (after [87]), where pilgrims moved from *left to right*. *Dark areas* correspond to phases of motion, *light colors* to stop phases. **c** Illustration of the “shell model”, in particular of situations where several pedestrians compete for the same gap, which causes coordination problems. **d** Stop-and-go waves resulting from the alternation of forward pedestrian motion and backward gap propagation

Herein,  $v_{\alpha}^{\max}$  is the maximum desired velocity and  $v_{\alpha}^0(0)$  the initial one, corresponding to the expected velocity of leaving. The time-dependent parameter

$$n_{\alpha}(t) = 1 - \frac{\bar{v}_{\alpha}(t)}{v_{\alpha}^0(0)} \quad (17)$$

reflects the nervousness, where  $\bar{v}_{\alpha}(t)$  denotes the average speed into the desired direction of motion. Altogether, long waiting times increase the desired speed  $v_{\alpha}^0$  or driving force  $v_{\alpha}^0(t)e_{\alpha}^0/\tau$ , which can produce high densities and inefficient motion. This further increases the waiting times, and so on, so that this tragic feedback can eventually trig-



ger so high pressures that people are crushed or falling and trampled. It is, therefore, imperative, to have sufficiently wide exits and to prevent counterflows, when big crowds want to leave [46].

**Transition to Stop-and-Go Waves** Recent empirical studies of pilgrim flows in the area of Makkah, Saudi Arabia, have shown that intermittent flows occur not only when bottlenecks are obvious. On January 12, 2006, pronounced stop-and-go waves have been even observed upstream of the 44 m wide entrance to the Jamarat Bridge [87]. While the pilgrim flows were smooth and continuous (“laminar”) over many hours, at 11:53 am stop-and-go waves suddenly appeared and propagated over distances of more than 30 m (see Fig. 7). The sudden transition was related to a significant drop of the flow, i. e. with the onset of congestion [87]. Once the stop-and-go waves set in, they persisted over more than 20 min.

This phenomenon can be reproduced by a recent model based on two continuity equations, one for forward pedestrian motion and another one for backward gap propagation [91]. The model was derived from a “shell model” (see Fig. 7) and describes very well the observed alternation between backward gap propagation and forward pedestrian motion.

**Transition to “Crowd Turbulence”** On the same day, around 12:19, the density reached even higher values and the video recordings showed a sudden transition from stop-and-go waves to *irregular* flows (see Fig. 8). These irregular flows were characterized by random, unintended

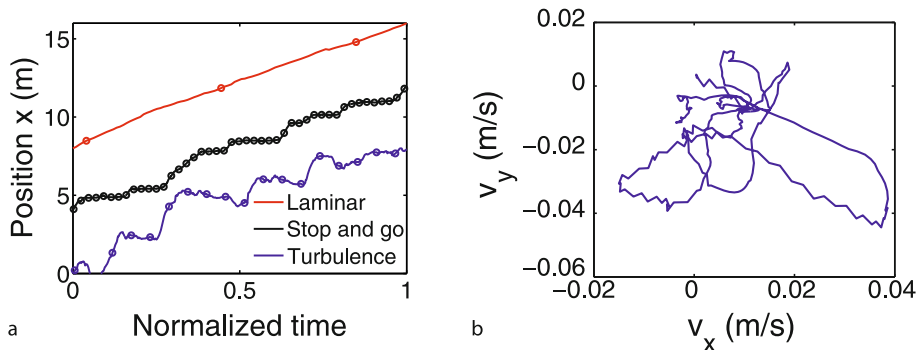
displacements into all possible directions, which pushed people around. With a certain likelihood, this caused them to stumble. As the people behind were moved by the crowd as well and could not stop, fallen individuals were trampled, if they did not get back on their feet quickly enough. Tragically, the area of trampled people grew more and more in the course of time, as the fallen pilgrims became obstacles for others [87]. The result was one of the biggest crowd disasters in the history of pilgrimage.

How can we understand this transition to irregular crowd motion? A closer look at video recordings of the crowd reveals that, at this time, people were so densely packed that they were moved involuntarily by the crowd. This is reflected by random displacements into all possible directions. To distinguish these irregular flows from laminar and stop-and-go flows and due to their visual appearance, we will refer to them as “*crowd turbulence*”.

As in certain kinds of fluid flows, “turbulence” in crowds results from a sequence of instabilities in the flow pattern. Additionally, one finds a sharply peaked probability density function of velocity increments

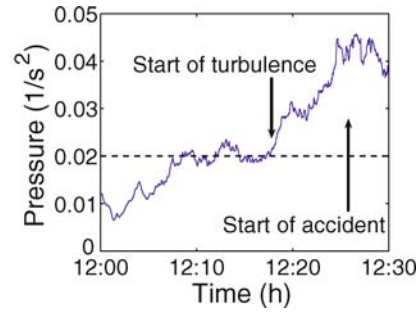
$$V_x^{\tau} = V_x(\mathbf{r}, t + \tau) - V_x(\mathbf{r}, t), \quad (18)$$

which is typical for turbulence [92], if the time shift  $\tau$  is small enough [87]. One also observes a power-law scaling of the displacements indicating self-similar behavior [87]. As large eddies are not detected, however, the similarity with *fluid* turbulence is limited, but there is still an analogy to turbulence at currency exchange markets [92]. Instead of vortex cascades like in turbulent fluids, one rather finds a hierarchical fragmentation dynamics: At extreme



**Pedestrian, Crowd and Evacuation Dynamics, Figure 8**

Pedestrian dynamics at different densities. **a** Representative trajectories (space-time plots) of pedestrians during the laminar, stop-and-go, and turbulent flow regime. Each trajectory extends over a range of 8 meters, while the time required for this stretch is normalized to 1. To indicate the different speeds, symbols are included in the curves every 5 seconds. While the laminar flow (*top line*) is fast and smooth, motion is temporarily interrupted in stop-and-go flow (*medium line*), and backward motion can occur in “turbulent” flows (*bottom line*). **b** Example of the temporal evolution of the velocity components  $v_x(t)$  into the average direction of motion and  $v_y(t)$  perpendicular to it in “turbulent flow”, which occurs when the crowd density is extreme. One can clearly see the irregular motion into all possible directions characterizing “crowd turbulence”. For details see [87]



**Pedestrian, Crowd and Evacuation Dynamics, Figure 9**

**Left:** Snapshot of the on-line visualization of “crowd pressure”. Red colors (see the lower ellipses) indicate areas of critical crowd conditions. In fact, the sad crowd disaster during the Muslim pilgrimage on January 12, 2006, started in this area. **Right:** The “crowd pressure” is a quantitative measure of the onset of “crowd turbulence”. The crowd disaster started when the “crowd pressure” reached particularly high values

densities, individual motion is replaced by mass motion, but there is a stick-slip instability which leads to “rupture” when the stress in the crowd becomes too large. That is, the mass splits up into clusters of different sizes with strong velocity correlations *inside* and distance-dependent correlations *between* the clusters.

“Crowd turbulence” has further specific features [87]. Due to the physical contacts among people in extremely dense crowds, we expect commonalities with granular media. In fact, dense driven granular media may form density waves, while moving forward [93], and can display turbulent-like states [94,95]. Moreover, under quasi-static conditions [94], force chains [96] are building up, causing strong variations in the strengths and directions of local forces. As in earthquakes [97,98] this can lead to events of sudden, uncontrollable stress release with power-law distributed displacements. Such a power-law has also been discovered by video-based crowd analysis [87].

### Some Warning Signs of Critical Crowd Conditions

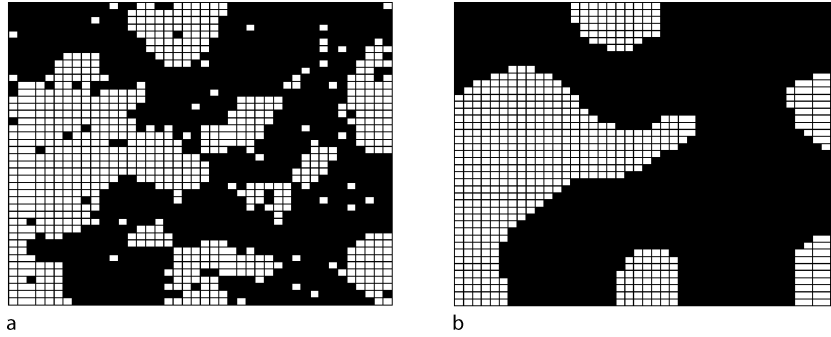
Turbulent waves are experienced in dozens of crowd-intensive events each year all over the world [88]. Therefore, it is necessary to understand why, where and when potentially critical situations occur. Viewing real-time video recordings is not very suited to identify critical crowd conditions: While the average density rarely exceeds values of 6 persons per square meter, the local densities can reach almost twice as large values [87]. It has been found, however, that even evaluating the local densities is not enough to identify the critical times and locations precisely, which also applies to an analysis of the velocity field [87]. The decisive quantity is rather the “crowd pressure”, i. e. the density, multiplied with the variance of speeds. It allows one to identify critical locations and times (see Fig. 9).

There are even advance warning signs of critical crowd conditions: The crowd accident on January 12, 2006 started about 10 minutes after “turbulent” crowd motion set in, i. e. after the “pressure” exceeded a value of  $0.02/\text{s}^2$  (see Fig. 9). Moreover, it occurred more than 30 min after stop-and-go waves set in, which can be easily detected in accelerated surveillance videos. Such advance warning signs of critical crowd conditions can be evaluated on-line by an automated video analysis system. In many cases, this can help one to gain time for corrective measures like flow control, pressure-relief strategies, or the separation of crowds into blocks to stop the propagation of shock-waves [87]. Such anticipative crowd control could increase the level of safety during future mass events.

### Evolutionary Optimization of Pedestrian Facilities

Having understood some of the main factors causing crowd disasters, it is interesting to ask how pedestrian facilities can be designed in a way that maximizes the efficiency of pedestrian flows and the level of safety. One of the major goals during mass events must be to avoid extreme densities. These often result from the onset of congestion at bottlenecks, which is a consequence of the breakdown of free flow and causes an increasing degree of compression. When a certain critical density is increased (which depends on the size distribution of people), this potentially implies high pressures in the crowd, particularly if people are impatient due to long delays or panic.

The danger of an onset of congestion can be minimized by avoiding bottlenecks. Notice, however, that jamming can also occur at widenings of escape routes [46]. This surprising fact results from disturbances due to pedestrians, who try to overtake each other and expand in the wider area because of their repulsive interactions.



**Pedestrian, Crowd and Evacuation Dynamics, Figure 10**

The evolutionary optimization based on Boolean grids [99] uses a two-stage algorithm. **a** In the randomization stage, obstacles are distributed over the grid with some randomness, thereby allowing for the generation of new topologies. **b** In the agglomeration stage, small nearby obstacles are clustered to form larger objects with smooth boundaries

These squeeze into the main stream again at the end of the widening, which acts like a bottleneck and leads to jamming. The corresponding drop of efficiency  $E$  is more pronounced,

1. if the corridor is narrow,
2. if the pedestrians have different or high desired velocities, and
3. if the pedestrian density in the corridor is high.

Obviously, the emerging pedestrian flows decisively depend on the geometry of the boundaries. They can be simulated on a computer already in the planning phase of pedestrian facilities. Their configuration and shape can be systematically varied, e. g. by means of evolutionary algorithms [28,100] and evaluated on the basis of particular mathematical performance measures [7]. Apart from the *efficiency*

$$E = \frac{1}{N} \sum_{\alpha} \frac{\mathbf{v}_{\alpha} \cdot \mathbf{e}_{\alpha}^0}{v_{\alpha}^0} \quad (19)$$

we can, for example, define the *measure of comfort*  $C = (1 - D)$  via the discomfort

$$D = \frac{1}{N} \sum_{\alpha} \frac{(\mathbf{v}_{\alpha} - \overline{\mathbf{v}_{\alpha}})^2}{(\mathbf{v}_{\alpha})^2} = \frac{1}{N} \sum_{\alpha} \left( 1 - \frac{\overline{\mathbf{v}_{\alpha}^2}}{(\mathbf{v}_{\alpha})^2} \right). \quad (20)$$

The latter is again between 0 and 1 and reflects the frequency and degree of sudden velocity changes, i. e. the level of discontinuity of walking due to necessary avoidance maneuvers. Hence, the optimal configuration regarding the pedestrian requirements is the one with the highest values of efficiency and comfort.

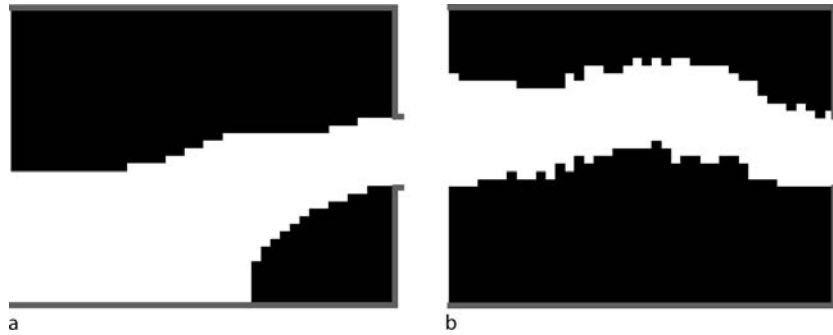
During the optimization procedure, some or all of the following can be varied:

1. the location and form of planned buildings,
2. the arrangement of walkways, entrances, exits, staircases, elevators, escalators, and corridors,
3. the shape of rooms, corridors, entrances, and exits,
4. the function and time schedule. (Recreation rooms or restaurants are often continuously frequented, rooms for conferences or special events are mainly visited and left at peak periods, exhibition rooms or rooms for festivities require additional space for people standing around, and some areas are claimed by queues or through traffic.)

In contrast to early evolutionary optimization methods, recent approaches allow to change not only the dimensions of the different elements of pedestrian facilities, but also to vary their topology. The procedure of such algorithms is illustrated in Fig. 10. Highly performing designs are illustrated in Fig. 11. It turns out that, for an emergency evacuation route, it is favorable if the crowd does not move completely straight towards a bottleneck. For example, a zigzag design of the evacuation route can reduce the pressure on the crowd upstream of a bottleneck (see Fig. 12). The proposed evolutionary optimization procedure can, of course, not only be applied to the design of new pedestrian facilities, but also to a reduction of existing bottlenecks, when suitable modifications are implemented.

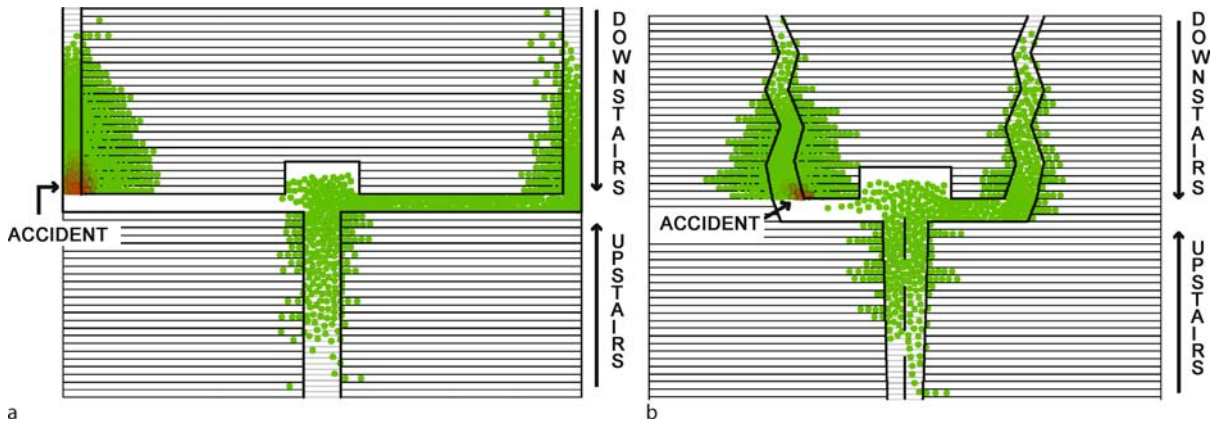
### Future Directions

In this contribution, we have presented a multi-agent approach to pedestrian and crowd dynamics. Despite the great effort required, pedestrian interactions can be well quantified by video tracking. Compared to other social interactions they turn out to be quite simple. Neverthe-



Pedestrian, Crowd and Evacuation Dynamics, Figure 11

Two examples of improved designs for cases with a bottleneck along the escape route of a large crowd, obtained with an evolutionary algorithm based on Boolean grids. People were assumed to move from left to right only. **a** Funnel-shaped escape route. **b** Zig-zag design



Pedestrian, Crowd and Evacuation Dynamics, Figure 12

**a** Conventional design of a stadium exit in an emergency scenario, where we assume that some pedestrians have fallen at the end of the downwards staircase to the left. The dark color indicates high pressures, since pedestrians are impatient and pushing from behind. **b** In the improved design, the increasing diameter of corridors can reduce waiting times and impatience (even with the same number of seats), thereby accelerating evacuation. Moreover, the zigzag design of the downwards staircases changes the pushing direction in the crowd. (After [8])

less, they cause a surprisingly large variety of self-organized patterns and short-lived social phenomena, where coordination or cooperation emerges spontaneously. For this reason, they are interesting to study, particularly as one can expect new insights into coordination mechanisms of social beings beyond the scope of classical game theory. Examples for observed self-organization phenomena in normal situations are lane formation, stripe formation, oscillations and intermittent clogging effects at bottlenecks, and the evolution of behavioral conventions (such as the preference of the right-hand side in continental Europe). Under extreme conditions (high densities or panic), however, coordination may break down, giving rise to “freezing-by-heating” or “faster-is-slower effects”, stop-and-go waves or “crowd turbulence”.

Similar observations as in pedestrian crowds are made in other social systems and settings. Therefore, we expect that realistic models of pedestrian dynamics will also promote the understanding of opinion formation and other kinds of collective behaviors. The hope is that, based on the discovered elementary mechanisms of emergence and self-organization, one can eventually also obtain a better understanding of the constituting principles of more complex social systems. At least the same underlying factors are found in many social systems: Non-linear interactions of individuals, time-dependence, heterogeneity, stochasticity, competition for scarce resources (here: Space and time), decision-making, and learning. Future work will certainly also address issues of perception, anticipation, and communication.



## Acknowledgments

The authors are grateful for partial financial support by the German Research Foundation (research projects He 2789/7-1, 8-1) and by the “Cooperative Center for Communication Networks Data Analysis”, a NAP project sponsored by the Hungarian National Office of Research and Technology under grant No. KCKHA005.

## Bibliography

### Primary Literature

- Hankin BD, Wright RA (1958) Passenger flow in subways. *Operat Res Q* 9:81–88
- Older SJ (1968) Movement of pedestrians on footways in shopping streets. *Traffic Eng Control* 10:160–163
- Weidmann U (1993) Transporttechnik der Fußgänger. In: *Schriftenreihe des Instituts für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau*. Institut für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau, Zürich
- Fruin JJ (1971) Designing for pedestrians: A level-of-service concept. In: *Highway research record*, Number 355: Pedestrians. Highway Research Board, Washington DC, pp 1–15
- Pauls J (1984) The movement of people in buildings and design solutions for means of egress. *Fire Technol* 20:27–47
- Whyte WH (1988) *City. Rediscovering the center*. Doubleday, New York
- Helbing D (1997) *Verkehrsdynamik*. Springer, Berlin
- Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transport Sci* 39(1):1–24
- Predtetschenski WM, Milinski AI (1971) Personenströme in Gebäuden – Berechnungsmethoden für die Projektierung. Müller, Köln-Braunsfeld
- Transportation Research Board (1985) *Highway Capacity Manual*, Special Report 209. Transportation Research Board, Washington DC
- Yuhaski SJ Jr, Macgregor Smith JM (1989) Modelling circulation systems in buildings using state dependent queueing models. *Queueing Syst* 4:319–338
- Garbrecht D (1973) Describing pedestrian and car trips by transition matrices. *Traffic Q* 27:89–109
- Ashford N, O’Leary M, McGinity PD (1976) Stochastic modelling of passenger and baggage flows through an airport terminal. *Traffic Engin Control* 17:207–210
- Borgers A, Timmermans H (1986) City centre entry points, store location patterns and pedestrian route choice behaviour: A microlevel simulation model. *Socio-Econ Plan Sci* 20:25–31
- Helbing D (1993) *Stochastische Methoden, nichtlineare Dynamik und quantitative Modelle sozialer Prozesse*. Ph.D. thesis University of Stuttgart, 1992 (published by Shaker, Aachen)
- Helbing D, Isobe M, Nagatani T, Takimoto K (2003) Lattice gas simulation of experimentally studied evacuation dynamics. *Phys Rev E* 67:067101
- Daamen W, Hoogendoorn SP (2003) Experimental research on pedestrian walking behavior (CDROM). In: *Proceedings of the 82nd annual meeting at the transportation research board*, Washington DC
- Isobe M, Helbing D, Nagatani T (2004) Experiment, theory, and simulation of the evacuation of a room without visibility. *Phys Rev E* 69:066132
- Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. *J Stat Mech* P10002
- Kretz T, Wölki M, Schreckenberg M (2006) Characterizing correlations of flow oscillations at bottlenecks. *J Stat Mech* P02005
- Henderson LF (1974) On the fluid mechanics of human crowd motion. *Transp Res* 8:509–515
- Hughes RL (2002) A continuum theory for the flow of pedestrians. *Transp Res B* 36:507–535
- Helbing D (1992) A fluid-dynamic model for the movement of pedestrians. *Complex Syst* 6:391–415
- Hoogendoorn SP, Bovy PHL (2000) Gas-kinetic modelling and simulation of pedestrian flows. *Transp Res Rec* 1710:28–36
- Helbing D (1991) A mathematical model for the behavior of pedestrians. *Behav Sci* 36:298–310
- Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. *Phys Rev E* 51:4282–4286
- Gipps PG, Marksjö B (1985) A micro-simulation model for pedestrian flows. *Math Comp Simul* 27:95–105
- Bolay K (1998) *Nichtlineare Phänomene in einem fluid-dynamischen Verkehrsmodell*. Master’s thesis, University of Stuttgart
- Blue VJ, Adler JL (1998) Emergent fundamental pedestrian flows from cellular automata microsimulation. *Transp Res Rec* 1644:29–36
- Fukui M, Ishibashi Y (1999) Self-organized phase transitions in cellular automaton models for pedestrians. *J Phys Soc Japan* 68:2861–2863
- Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. *Physica A* 267:487–498
- Klüpfel H, Meyer-König M, Wahle J, Schreckenberg M (2000) Microscopic simulation of evacuation processes on passenger ships. In: Bandini S, Worsch T (eds) *Theory and practical issues on cellular automata*. Springer, London
- Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a 2-dimensional cellular automaton. *Physica A* 295:507–525
- Gopal S, Smith TR (1990) NAVIGATOR: An AI-based model of human way-finding in an urban environment. In: Fischer MM, Nijkamp P, Papageorgiou YY (eds) *Spatial choices and processes*. North-Holland, Amsterdam, pp 169–200
- Reynolds CW (1994) Evolution of corridor following behavior in a noisy world. In: Cliff D, Husbands P, Meyer J-A, Wilson S (eds) *From animals to animats 3: Proceedings of the third international conference on simulation of adaptive behavior*. MIT Press, Cambridge, pp 402–410
- Helbing D (1992) A mathematical model for attitude formation by pair interactions. *Behav Sci* 37:190–214
- Helbing D, Molnár P, Farkas I, Bolay K (2001) Self-organizing pedestrian movement. *Env Planning B* 28:361–383
- Klockgether J, Schwefel H-P (1970) Two-phase nozzle and hollow core jet experiments. In: Elliott DG (ed) *Proceedings of the eleventh symposium on engineering aspects of magnetohydrodynamics*. California Institute of Technology, Pasadena, pp 141–148

39. Helbing D (1992) A mathematical model for behavioral changes by pair interactions. In: Haag G, Mueller U, Troitzsch KG (eds) *Economic evolution and demographic change. Formal models in social sciences*. Springer, Berlin, pp 330–348
40. Miller NE (1944) Experimental studies of conflict. In: Mc Hunt VJ (ed) *Personality and the behavior disorders*, vol 1. Ronald, New York
41. Miller NE (1959) Liberalization of basic S-R-concepts: Extension to conflict behavior, motivation, and social learning. In: Koch S (ed) *Psychology: A study of science*, vol 2. McGraw Hill, New York
42. Lewin K (1951) *Field theory in social science*. Harper, New York
43. Helbing D (1994) A mathematical model for the behavior of individuals in a social field. *J Math Sociol* 19(3):189–219
44. Hoogendoorn S, Bovy PHL (2003) Simulation of pedestrian flows by optimal control and differential games. *Optim Control Appl Meth* 24(3):153–172
45. Johansson A, Helbing D, Shukla PK (2007) Specification of the social force pedestrian model by evolutionary adjustment to video tracking data. *Adv Complex Syst* 10:271–288
46. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. *Nature* 407:487–490
47. Kerridge J, Chamberlain T (2005) Collecting pedestrian trajectory data in real-time. In: Waldau N, Gattermann P, Knoflach H, Schreckenberg M (eds) *Pedestrian and evacuation dynamics '05*. Springer, Berlin
48. Hoogendoorn SP, Daamen W, Bovy PHL (2003) Extracting microscopic pedestrian characteristics from video data (CDROM). In: *Proceedings of the 82nd annual meeting at the transportation research board*. Mira Digital, Washington DC
49. Teknomo K (2002) Microscopic pedestrian flow characteristics: Development of an image processing data collection and simulation model. Ph D thesis, Tohoku University Japan
50. Kadanoff LP (1985) Simulating hydrodynamics: A pedestrian model. *J Stat Phys* 39:267–283
51. Stanley HE, Ostrowsky N (eds) (1986) *On growth and form*. Nijhoff, Boston
52. Arns T (1993) Video films of pedestrian crowds. Stuttgart
53. Stølum H-H (1996) River meandering as a self-organization process. *Nature* 271:1710–1713
54. Rodríguez-Iturbe I, Rinaldo A (1997) *Fractal river basins: Chance and self-organization*. Cambridge University, Cambridge
55. Helbing D, Farkas I, Vicsek T (2000) Freezing by heating in a driven mesoscopic system. *Phys Rev Lett* 84:1240–1243
56. Schelling T (1971) Dynamic models of segregation. *J Math Sociol* 1:143–186
57. Helbing D, Platkowski T (2000) Self-organization in space and induced by fluctuations. *Int J Chaos Theory Appl* 5(4):47–62
58. Ando K, Oto H, Aoki T (1988) Forecasting the flow of people. *Railw Res Rev* 45(8):8–13 (in Japanese)
59. Smith RA, Dickie JF (eds) (1993) *Engineering for crowd safety*. Elsevier, Amsterdam
60. Dräger KH, Løvås G, Wiklund J, Soma H, Duong D, Violas A, Lanérés V (1992) EVACSIM – A comprehensive evacuation simulation tool. In: *The proceedings of the 1992 Emergency Management and Engineering Conference*. Society for Computer Simulation, Orlando, pp 101–108
61. Ebihara M, Ohtsuki A, Iwaki H (1992) A model for simulating human behavior during emergency evacuation based on classificatory reasoning and certainty value handling. *Micro-comput Civ Engin* 7:63–71
62. Ketchell N, Cole S, Webber DM, Marriott CA, Stephens PJ, Brearley IR, Fraser J, Doherty J, Smart J (1993) The EGRESS code for human movement and behaviour in emergency evacuations. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 361–370
63. Okazaki S, Matsushita S (1993) A study of simulation model for pedestrian movement with evacuation and queuing. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 271–280
64. Still GK (1993) New computer system can predict human behaviour response to building fires. *Fire* 84:40–41
65. Still GK (2000) *Crowd dynamics*. Ph.D. thesis, University of Warwick
66. Thompson PA, Marchant EW (1993) Modelling techniques for evacuation. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 259–269
67. Løvås GG (1998) On the importance of building evacuation system components. *IEEE Trans Engin Manag* 45:181–191
68. Hamacher HW, Tjandra SA (2001) Mathematical modelling of evacuation problems: A state of the art. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin, pp 227–266
69. Keating JP (1982) The myth of panic. *Fire J* 57–61, 147
70. Elliott D, Smith D (1993) Football stadia disasters in the United Kingdom: Learning from tragedy? *Ind Env Crisis Q* 7(3):205–229
71. Jacobs BD, 't Hart P (1992) Disaster at Hillsborough Stadium: A comparative analysis. In: Parker DJ, Handmer JW (eds) *Hazard management and emergency planning*, Chapt 10. James and James Science, London
72. Canter D (ed) (1990) *Fires and human behaviour*. Fulton, London
73. Mintz A (1951) Non-adaptive group behavior. *J Abnorm Norm Soc Psychol* 46:150–159
74. Miller DL (1985) Introduction to collective behavior (Fig. 3.3 and Chap. 9). Wadsworth, Belmont
75. Coleman JS (1990) *Foundations of social theory*, Chaps. 9 and 33. Belknap, Cambridge
76. Johnson NR (1987) Panic at "The Who Concert Stampede": An empirical assessment. *Soc Probl* 34(4):362–373
77. LeBon G (1960) *The crowd*. Viking, New York
78. Quarantelli E (1957) The behavior of panic participants *Sociol Soc Res* 41:187–194
79. Smelser NJ (1963) *Theory of collective behavior*. Free Press, New York
80. Brown R (1965) *Social psychology*. Free Press, New York
81. Turner RH, Killian LM (1987) *Collective behavior*, 3rd edn. Prentice Hall, Englewood Cliffs
82. Bryan JL (1985) Convergence clusters. *Fire J* 27–30, 86–90
83. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211:1390–1396
84. Axelrod R, Dion D (1988) The further evolution of cooperation. *Science* 242:1385–1390
85. Glance NS, Huberman BA (1994) The dynamics of social dilemmas. *Scientific American* 270:76–81
86. Kelley HH, Condry JC Jr, Dahlke AE, Hill AH (1965) Collective behavior in a simulated panic situation. *J Exp Soc Psychol* 1:20–54

87. Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: An empirical study. *Phys Rev E* 75:046109
88. Fruin JJ (1993) The causes and prevention of crowd disasters. In: Smith RA, Dickie JF (eds) *Engineering for crowd safety*. Elsevier, Amsterdam, pp 99–108
89. Ristow GH, Herrmann HJ (1994) Density patterns in two-dimensional hoppers. *Phys Rev E* 50:R5–R8
90. Wolf DE, Grassberger P (eds) (1997) *Friction, arching, contact dynamics*. World Scientific, Singapore
91. Helbing D, Johansson A, Mathiesen J, Jensen HM, Hansen A (2006) Analytical approach to continuous and intermittent bottleneck flows. *Phys Rev Lett* 97:168001
92. Ghashghaie S, Breyman W, Peinke J, Talkner P, Dodge Y (1996) Turbulent cascades in foreign exchange markets. *Nature* 381:767–770
93. Peng G, Herrmann HJ (1994) Density waves of granular flow in a pipe using lattice-gas automata. *Phys Rev E* 49:R1796–R1799
94. Radjai F, Roux S (2002) Turbulentlike fluctuations in quasistatic flow of granular media. *Phys Rev Lett* 89:064302
95. Sreenivasan KR (1990) Turbulence and the tube. *Nature* 344:192–193
96. Cates ME, Wittmer JP, Bouchaud J-P, Claudin P (1998) Jamming, force chains, and fragile matter. *Phys Rev Lett* 81:1841–1844
97. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. *Phys Rev Lett* 88:178501
98. Johnson PA, Jia X (2005) Nonlinear dynamics, granular media and dynamic earthquake triggering. *Nature* 437:871–874
99. Johansson A, Helbing D (2007) Pedestrian flow optimization with a genetic algorithm based on Boolean grids. In: Waldau N, Gattermann P, Knoflach H, Schreckenberg M (eds) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin, pp 267–272
100. Baeck T (1996) *Evolutionary algorithms in theory and practice*. Oxford University Press, New York

## Books and Reviews

- Decicco PR (ed) (2001) *Evacuation from fires*. Baywood, Amityville
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Helbing D, Molnár P, Farkas I, Bolay K (2001) Self-organizing pedestrian movement. *Environ Plan B* 28:361–383
- Helbing D, Buzna L, Johansson A, Werner T (2005) *Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions*. *Transp Sci* 39(1):1–24
- Le Bon G (2002) *The Crowd*. Dover, New York (1st edn: 1895)
- Predtechenskii VM, Milinskii AI (1978) *Planning for foot traffic flow in buildings*. Amerind, New Delhi
- Schreckenberg M, Sharma SD (eds) (2002) *Pedestrian and evacuation dynamics*. Springer, Berlin
- Smith RA, Dickie JF (eds) (1993) *Engineering for crowd safety*. Elsevier, Amsterdam
- Still GK (2000) *Crowd Dynamics*. Ph.D thesis, University of Warwick
- Surowiecki J (2005) *The Wisdom of Crowds*. Anchor, New York
- Tubbs J, Meacham B (2007) *Egress design solutions: A guide to evacuation and crowd management planning*. Wiley, New York
- Waldau N, Gattermann P, Knoflach H (eds) (2006) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin

- Weidmann U (1993) *Transporttechnik der Fußgänger*. In: Schriftenreihe des Institut für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau 90. ETH Zürich

## Percolation in Complex Networks

REUVEN COHEN<sup>1</sup>, SHLOMO HAVLIN<sup>2</sup>

<sup>1</sup> Department of Mathematics, Bar-Ilan University, Ramat-Gan, Israel

<sup>2</sup> Department of Physics, Bar-Ilan University, Ramat-Gan, Israel

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Percolation Thresholds and Network Robustness](#)

[Epidemics and Immunization](#)

[The Generating Functions Method](#)

[Critical Exponents and Fractal Dimensions](#)

[Optimal Paths and Minimum Spanning Trees](#)

[Fragmentation of Social Networks](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Graph** A set of nodes (sites) and edges (links or bonds) connecting them.

**Weighted graph** A graph where each edge is assigned a (usually non-negative) weight.

**Random graph** A graph selected from an ensemble (probability space) of graphs.

**Degree** Number of edges emanating from a node.

**Scale free network** A network whose nodes' degrees are distributed according to a power law.

**Shortest path** The path with minimum number of edges connecting two nodes.

**Optimal path** In a weighted graph – the path with minimum total weight connecting two nodes.

**Loop** A path that start and ends at the same node.

**Tree** A connected graph (a graph consisting of a single component) with no loops.

**Minimum spanning tree** In a weighted graph – the tree subgraph of the graph with the minimum total weight.

**Component** The set of nodes reachable from a given node. The nodes of a component are all reachable from each other.

**Giant component** The component of a graph with size (number of nodes) of order of the number of nodes in the graph.

**Percolation theory** The theory studying the connectivity behavior of networks when a fraction of the nodes or edges are removed. Site (or node) percolation involves occupying a fraction,  $p$ , of the nodes of the graph, or alternatively, removing a fraction  $q = 1 - p$ . In bond (or edge) percolation edges are occupied, or removed, with some probability. A combined site-bond percolation, where both processes occur simultaneously, is also considered.

**Percolation threshold** The fraction,  $p_c$  of occupied nodes or edges, under the graph is fragmented into small components, and above which a giant component emerges.

### Definition of the Subject

In this chapter we survey the application of percolation theory to several random network classes, and in particular, to scale free networks. We show how ideas from percolation theory can be applied to the study of robustness and vulnerability of random networks. We show how percolation techniques can be applied also to understand phenomena such as immunization and epidemic spreading in populations and computer networks, minimum spanning trees and communication paths and fragmentation in social networks.

### Introduction

In recent years considerable interest has been given to real world networks. The importance of technological networks such as the Internet and WWW, as well as the availability of large scale data sets on social, biological and technological networks made this subject approachable and popular.

The main model used in the study of complex networks was Erdős–Rényi (ER) random graphs [19,20,21] (also presented earlier by Rapaport [33]), which are graphs having  $N$  nodes (that is, sites or entities) and  $M$  edges (links, or connections between the nodes) distributed randomly between them, or alternatively, the almost identical model, having every pair of nodes connect with a constant probability. One of the key discoveries in recent years was that many real world networks, including the Internet, WWW and many biological and social networks, are not described well by the ER model [1,18,28,32]. One of the first deviations from the model to be noticed was the tendency of many real world networks to have high clustering [28], i. e., neighbors of the same node tend to connect

between them (i. e., share an edge). This discovery has led to the presentation of the small world model [37].

The small world model is based on some lattice, such as a one-dimensional ring or a higher dimensional grid, in which rewiring occurs. Some (usually small) fraction,  $\varphi$  of the links in the lattice are removed, and instead, new links are added randomly between the nodes. When  $\varphi$  is moderately small the generated graphs have the desired properties of high clustering, while the average distance between nodes is small (of the order  $\log N$ ), as in random graphs (as opposed to  $N^{1/d}$  as in a  $d$ -dimensional grid).

The second deviation to be noticed was the deviation of the degree sequence from the expectation of random graph theory. The degree of a node is the number of links (or edges) emanating from it, i. e., the number of neighbors it has in the graph. The number of nodes of degree  $k$  in a graph will be denoted  $n(k)$ , and the degree distribution, i. e., the probability that a random node has degree  $k$ , is  $P(k) = n(k)/N$ . In an ER random graph the expected degree sequence is a Poisson distribution [5],  $P(k) = e^{-C} C^k/k!$ , where  $P(k)$  is the probability of a node to have degree  $k$ , and  $c$  is the average degree ( $C = 2M/N$ ). In many real world networks, including the Internet and WWW, it was observed that the degree distribution is actually a power law,

$$P(k) = ck^{-\gamma}, \quad m \leq k \leq U, \quad (1)$$

where  $c$  is a normalization factor,  $m$  and  $U$  are the minimum and maximum degrees, respectively, and  $\gamma$  is some exponent characterizing the distribution. In most networks studied  $\gamma$  has been found to lay in the range  $2 < \gamma < 3$  [1,18,28,32]. These networks have become known as scale free networks, due to the lack of typical scale for the degree distribution. It should be noted that while  $m$  must be supplied externally for the distribution to be normalizable,  $U$  can be omitted, and will be determined naturally as the extreme value statistics of  $N$  variables, which gives  $U \sim mN^{1/(\gamma-1)}$  in this case [13].

Several models have been developed for the understanding and study of scale free networks. The question of the reason for these networks' formation has been addressed by the Barabási–Albert model [3]. Many variants on the model have been studied since (see, e. g., [23]). Here we will focus on the configuration model, or the generalized random graph model [4], which is an equilibrium model for random graphs with a given degree distribution, producing all graphs having a given degree distribution with uniform probability. The model starts by having  $N$  distinct nodes and randomly selecting the degree of each of these nodes from the given degree distribution. Each node,  $i$  is then fitted with  $k_i$  “stubs”, where  $k_i$  is its degree, drawn



from the distribution  $P(k)$ . After all nodes' degrees have been selected, a random matching of the stubs is selected, by choosing random pairs of stubs and pairing them (i. e., connecting the nodes by an edge and removing both stubs) until no stubs are left. In some cases a single stub is left, and also there may be edges connecting a node to itself or more than one edge connecting a given pair of nodes. These cases can be safely ignored as they only have a small effect on the final graphs obtained.

### Percolation Thresholds and Network Robustness

One of the fundamental questions regarding a network's structure is its connectivity properties, i. e., what are the properties of the distinct components (or clusters) of the network. A component of a graph is the set of nodes reachable from a single node by following edges in the graph. Notice that in an (undirected) graph, the property of path connectedness is symmetric and transitive, i. e., if a node  $a$  is reachable from node  $b$ , then the inverse path also exists, and if  $c$  is reachable from  $a$  it is also reachable from  $b$ . Therefore, a component is uniquely defined by any node belonging to it. A network is said to be connected if all nodes in it belong to a single component, that is, each node is reachable from each other node.

Random graphs are locally tree-like, i. e., the number of closing a loop for a set of less than order  $N$  nodes is negligible. This also implies that below the percolation threshold, where all components (clusters) are small, almost all components are trees, i. e., possess of no loops.

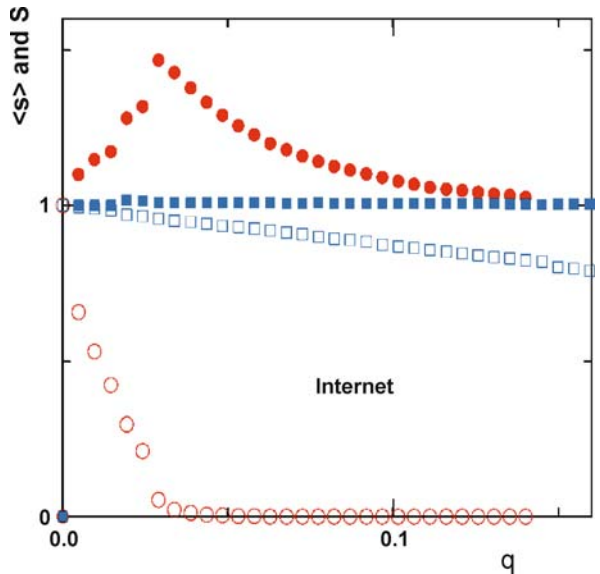
To determine the percolation threshold it should be noticed that at the critical point every node reached by following a link from a previously visited node should have, on average, exactly one more link through which new nodes can be reached. If the average number of outgoing links is less than one, the uncovering process of the component will quickly decay, and only small components will be present. If the average number of outgoing links is larger than one, the size of the largest component will be proportional to that of the entire graph, i. e., a giant component, of size  $O(N)$  will exist. This may lead to the conclusion that the average degree,  $\langle k \rangle$ , needs be two or more for a giant component to exist. However, the node reached by following a link is not chosen uniformly. The probability of reaching a node by following a link is proportional to its degree,  $P_i = k_i / (N \langle k \rangle)$ . The average outgoing degree of a node reached by following a link is therefore,  $\sum (k-1)n(k)P_i = \langle k(k-1) \rangle / \langle k \rangle = \kappa - 1$ , where  $\kappa = \langle k^2 \rangle / \langle k \rangle$  is the ratio of the first two moments of the degree distribution. It is this quantity that should be compared to one to determine whether a giant compo-

nent exists [13,26]. Therefore, a giant component exists if and only if  $\kappa > 2$ .

In a percolation setting the nodes or edges are removed with probability  $q$ , or, alternatively, occupied with probability  $p = 1 - q$ . This model may represent random failures of nodes in the network, such as random failures of routers or links in the Internet [2]. The average number of outgoing links should be multiplied then by the probability,  $p_b$ , that the link is occupied, and by  $p_s$ , the probability that the node reached from the link is occupied, or not deleted. Therefore, the condition for the existence of a giant component becomes  $p_b p_s (\kappa - 1) > 1$ . Alternatively,  $p_c$ , the critical site or bond percolation threshold is given by [13]

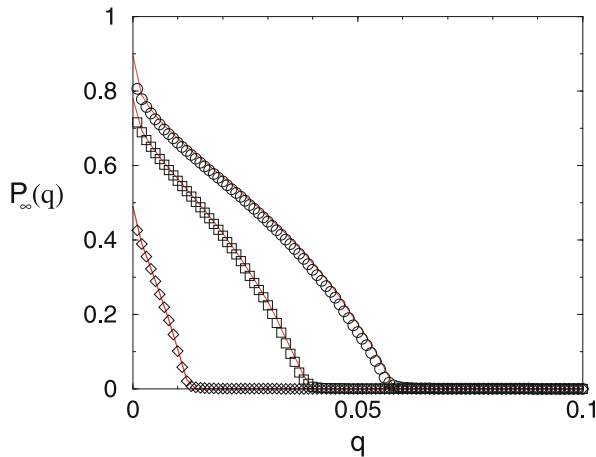
$$p_c = \frac{1}{\kappa - 1}. \quad (2)$$

Notice that the critical threshold depends only on the first two moments of the degree distribution. Furthermore, in a scale free network with  $\gamma \leq 3$  the second moment of the degree distribution diverges in the limit of infinite network size. Therefore, the critical threshold for this class of networks approaches 0, indicating that these networks are resilient to any finite fraction of random node failures [13]. Figure 1 presents the results of random node failure on the Internet as compared to an ER network.



Percolation in Complex Networks, Figure 1

Results of targeted removal of a fraction  $q$  of the nodes from an ER graph (circles) and a partial Internet view (squares). Full symbols represent the size of the second largest component and empty symbols represent  $P_\infty$ , the relative size of the largest component. After [2]



**Percolation in Complex Networks, Figure 2**

The relative size of the largest component,  $P_\infty$ , as a function of the fraction of targeted removed nodes, for scale free networks with  $m = 1$  and  $\gamma = 2.5$  (circles),  $\gamma = 2.8$  (squares), and  $\gamma = 3.3$  (diamonds). After [14]

In case the node removal is not random this situation may change drastically. The most well studied case is that of removal of the highest degree nodes, modeling an intentional, targeted attack on the most important nodes in the network. In this case calculations similar to the above lead to the conclusion that the percolation threshold is finite and small [2,9,14]. Figure 2 illustrates the results of targeted removal of a fraction  $q$  in scale free networks.

### Epidemics and Immunization

The spread of epidemics in a population can be modeled as a dynamical process in a network. Each node represents an individual and the links represent interaction between individuals that allows transmission of the epidemic. Several models exist for epidemic transmission, depending on the type of epidemic. The most commonly used models are the Susceptible-Infected-Susceptible (SIS) model and the Susceptible-Infected-Removed (SIR) model. In both models it is assumed that nodes in the susceptible state are susceptible to the epidemic, i.e., may be infected when they come in contact with an infected individual. In the infected state the individual is infected by the epidemic and may infect other individuals, and in the removed (or recovered) state the individual is no more infected or infective and also is no longer susceptible to the disease. This state may occur due to recovery from the disease while the individual remains immunized against the disease or due to death of the infected individual. The SIS model assumes that recovered individuals are again susceptible and the SIR model

assumes that each individual may only be infected once in a lifetime.

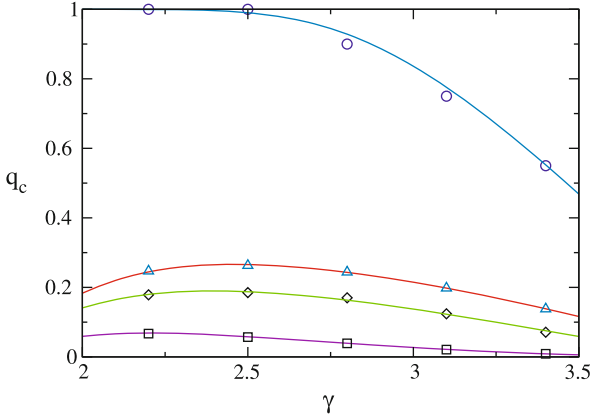
In the SIS model an epidemic can either quickly decay and vanish or prevail for a long period, during which a large (finite) fraction of the population is infected. In the SIS model it was shown that in scale free networks with  $\gamma \leq 3$  the epidemic always prevails, regardless of the infection rate [30].

Consider epidemic spreading in the SIR model. Assume each infected node infects each of its neighbors with rate  $R$ , and has a constant infection time of  $T$ . For each neighbor, the probability of the neighbor being infected by an infected node is  $p = 1 - e^{-RT}$ . This can be viewed as the probability of the edge between the two nodes is occupied, i.e., can actually be used for transmission of the epidemic. Therefore, the SIR model can be mapped to a bond percolation model (Notice that in non homogeneous cases of node dependent rates a more complicated model is needed. See [27].)

The SIR model can therefore be solved by solving the bond percolation problem in the network [22,27,34]. Every edge is occupied with probability  $p$ , and the epidemic can reach an endemic state (i.e., infect  $O(N)$  nodes with finite probability) if  $p > p_c$ , and will quickly decay, infecting only a negligible portion of the population if  $p < p_c$ . Furthermore, the distribution of the size of the epidemic outbreak is determined by the sizes of the graph components. If a giant component exists, the probability of a single infected individual to induce an endemic state of the population is  $P_\infty$ , the size of the giant component, and the size of the outbreak is the size of the component to which this individual belongs.

To prevent epidemic outbreaks it is usually desirable to immunize the population and thus prevent the epidemic. In many cases it is difficult to immunize the entire population, and only a fraction is immunized. Each immunized individual is no longer susceptible to the disease, and can be viewed as removed from the network. The immunization process can therefore be viewed as a site percolation process, where each node is removed with probability  $q$ , or occupied with probability  $p = 1 - q$ . The epidemic progression can then be mapped into a site-bond percolation problem.

In order for the immunization to be highly efficient, it is desirable to surpass the percolation threshold in the immunization process, to ensure that the epidemic can not reach an endemic state. Since, as stated above, randomly immunizing a fraction of the population can be a highly inefficient process, requiring immunization of nearly 100% of the population, it has been suggested that a more efficient method for immunization is devised. The



**Percolation in Complex Networks, Figure 3**

Critical immunization threshold,  $q_c$ , as a function of  $\gamma$  in scale-free networks (with  $m = 1$ ), for the random immunization ( $\circ$ ), acquaintance immunization ( $\Delta$ ), double acquaintance immunization ( $\diamond$ ), and targeted immunization ( $\square$ ) strategies. Curves represent analytical results, while data points represent simulation data, for a population  $N = 10^6$  (Due to the population's final size  $q_c < 1$  for random immunization even when  $\gamma < 3$ ). After [16]

simplest such method involves the immunization of the highest degree nodes in the population [31]. In this case immunizing a population will require vaccinating only a finite, and relatively small, fraction of the population.

In case only partial knowledge of the population exists it is sometimes also possible to immunize the population efficiently [17]. However, a different method, requiring no global knowledge exists. In this method, “acquaintance immunization”, a fraction of randomly selected individuals are requested to point to one of their contacts, also randomly selected. The pointed contacts are then immunized. Although this is a seemingly random process a node having a high degree is immunized with very high probability, and the process behaves effectively as a targeted immunization of high degree nodes. See [16] for analytical treatment and Fig. 3 for illustration of the various immunization thresholds.

### The Generating Functions Method

To allow the calculation of different properties of random networks it was proposed in [9,29] to use the generating function formalism (see, e. g., [38]). In this formalism a list of numbers  $A_i$  is treated as the coefficients of a formal power series  $A(x) = \sum_i A_i x^i$ . This treatment simplifies many equations regarding the variables  $A_i$ , and, in many cases, simplifies the calculation of the asymptotic behavior of the coefficients for large  $i$ .

In [9,29] a power series is built for the degree distribution,

$$G_0(x) = \sum_k P(k)x^k, \quad (3)$$

and for the distribution of outgoing links from a node reached by following a link,

$$G_1(x) = \sum_k \frac{kP(k)}{\langle k \rangle} x^{k-1} = \frac{G'_0(x)}{G'_0(1)}, \quad (4)$$

where  $G'_0 = dG_0/dx$ .

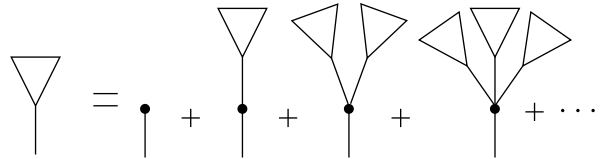
A branch in the network is a link, traversed in one direction, and all the nodes reachable by following this link in this direction. This includes the node reached by following this link, and all the nodes reached by the branches emanating from the outgoing links of this node. An illustration of this recursive definition is in Fig. 4. A branch may be either finite or infinite, in which case, for a finite graph, it will reach  $O(N)$  nodes and have many loops (in which case the branch description is no longer useful). A link and the node reached by following it are occupied with probability  $p_b p_s$ , and the generating function for the number of descendants of the reached node is  $G_1$ . Each of the descendants is a new branch. Therefore, the generating function for the size of a branch is given by

$$H_1(x) = 1 - p_s p_b + p_s p_b G_1(H_1(x)). \quad (5)$$

A component in the graph is a node, and all the nodes reachable from it. Each of the links of a node leads to a branch. The generating function for the degree of a node is  $G_0$ , and the probability it is occupied is  $p_s$ . Therefore, the generating function for the component size distribution is,

$$H_0(x) = 1 - p_s + p_s G_0(H_1(x)), \quad (6)$$

where  $H_1$  is determined by Eq. (5). Again, as above, a component may be finite or infinite. Since  $H_0$  is the generating function for the finite cluster distribution, its normalization  $H_0(1)$  is the total probability that a node belongs to a finite component, and therefore its complement



**Percolation in Complex Networks, Figure 4**

Illustration of the structure of a branch. A branch can contain either a link leading to a node with no outgoing links, or to a node having one or more outgoing links, each leading to another branch. After [29]

$1 - H_0(1)$  is the probability that a node belongs to the giant component. Thus, the size of the giant component (that is, the fraction of nodes belonging to the giant component) is given by

$$P_\infty = 1 - H_0(1) = 1 - p_s + p_s G_0(u), \quad (7)$$

and  $u = H_1(1)$  is given by the self consistent equation derived from Eq. (5)

$$u = 1 - p_s p_b + p_s p_b G_1(u). \quad (8)$$

### Critical Exponents and Fractal Dimensions

Percolation problems, as well as other critical phenomena, are known to present a universal behavior near and at the critical point. That means that many properties of the critical structure behave as power laws and that near the critical point many sizes behave as powers of  $p - p_c$ . The universality is pronounced by the fact that the exponents in the different power laws do not depend on the microscopic details of the problem, but only on the large scale details, in particular, the dimension of the space and the symmetries. In percolation, for example, it is known that slightly above the critical point the size of the giant component behaves as  $P_\infty \sim (p - p_c)^\beta$  and that the number of components of size  $s$  at criticality decays as  $n(s) \sim s^{-\tau}$ . Both  $\beta$  and  $\tau$  depend only on the dimensionality and not on the microscopic details (e. g., are the same for two dimensional square and triangular lattices, etc.).

Networks can be considered infinite dimensional objects. As mentioned above, the number of nodes at a distance (number of hops) at most  $\ell$  from a node behaves as  $A^\ell$  for some  $A > 1$ . For large values of  $\ell$  this is larger than  $\ell^d$ , obtained for any finite dimension  $d$ . This property also leads to the impossibility of embedding networks in any finite dimension.

For percolation theory it is known [8,35] that the upper critical dimension is 6, i. e., percolation on grids of any dimension larger than 6 behave similarly to percolation in infinite dimension (also known as mean field percolation, and usually studied using the Cayley tree model [8]). In mean field percolation it is known that  $\beta = 1$  and  $\tau = 2.5$ . For dimensions less than 6 it is known that  $\beta < 1$  and that  $\tau < 2.5$ .

To determine the critical exponents for random networks one can use the generating function formalism presented above. The size of the giant component is given by Eq. (7), where the value of  $u$  in this equation is the solution of Eq. (8). At and below the critical concentration,  $p_c$ , the size of the giant component is  $P_\infty = 0$ , since there are only finite components. This corresponds to a solution in

which  $u = 1$  and  $H_0(1) = 1$ . At  $p = p_c + \delta$  it is expected that  $u$  is close to 1, i. e.,  $u = 1 - \epsilon$ . Substituting this into Eq. (8) yields

$$1 - \epsilon = 1 - p_c - \delta + (p_c + \delta)G_1(1 - \epsilon). \quad (9)$$

Expanding  $G_1$  into a power series yields

$$\epsilon = (p_c + \delta)(1 - G_1(1) - G_1'(1)\epsilon - G_1''(1)\epsilon^2/2 - \dots). \quad (10)$$

This equation is self consistent and gives a non trivial solution only when  $p_c = G_1'(1) = 1/(\kappa - 1)$  as obtained above. The solution obtained then is  $\epsilon \propto \delta$ . The solution of Eq. (7) then is  $P_\infty \propto \delta = (p - p_c)$ , similar to infinite dimensional percolation. Using similar expansions of Eqs. (7) and (8), now at the critical point,  $p = p_c$  it can be shown that the probability of a node to belong to a component of size  $s$  is proportional to,  $p_s \propto s^{-3/2}$ , implying that the number of components of size  $s$  behaves as  $n(s) \propto s^{-5/2}$ . Both these exponents are the same as obtained for infinite dimension percolation.

The above treatment is correct, however, only assuming the sums in Eqs. (7) and (8) can be expanded in a power series. This is true only if the degree distribution,  $P(k)$  decays quickly enough, say exponentially. In the case of a power law degree distribution, the series expansion is incorrect, and one needs to resort to other methods of obtaining the asymptotic behavior. The main mechanism for obtaining such asymptotics is using Abelian and Tauberian theorems. These theorems relate the decay of the coefficients of a power series and its behavior near a singular point of the function in the complex plane. Using these methods it can be shown [12] that the behavior of Eq. (8) near criticality becomes

$$1 - \epsilon = 1 - p_c - \delta + (p_c + \delta)(G_1(1) + G_1'(1)\epsilon + G_1''(1)\epsilon^2/2 + \dots + C\epsilon^{\gamma-1} + \dots). \quad (11)$$

Thus, when  $\gamma < 3$  the nonanalytic term  $C\epsilon^{\gamma-1}$  dominates the linear term, and for  $3 < \gamma < 4$  it dominates the quadratic term.

In the most interesting case of  $3 < \gamma < 4$  the percolation threshold is finite, as seen from Eq. (2). However, using the expansion in Eq. (11) it can be seen that near the critical point  $P_\infty \propto (p - p_c)^{1/(\gamma-3)}$ , so  $\beta = 1/(\gamma - 3)$ . Similarly, it can be shown that  $\tau = (2\gamma - 3)/(\gamma - 2)$  in this regime. Both exponents return to their mean field value for  $\gamma > 4$ .

Another common characteristic of critical phenomena is the fractal behavior at the critical point. For high dimensional percolation at the critical dimension  $d_c = 6$  and above it is known that the fractal dimension of the



largest components is  $d_f = 4$ . This implies that the size of the largest component behaves as  $S \sim L^4$ , where  $L$  is the linear dimension of the grid. Since the total grid size is  $N \sim L^6$  it follows that  $S \sim N^{2/3}$ . Also, it is known that each branch of the components at criticality behaves as an independent random walk. Since random walks have  $\ell \sim L^2$ , where  $\ell$  is the number of random walk steps, representing the distance on the fractal itself, also known as the chemical distance. This implies that  $S \sim \ell^2$ . This dimension,  $d_f = 2$  is known as the chemical dimension [23], and since in a network no embedding space is present, and therefore  $L$  is not well defined, the chemical dimension is the most appropriate measure to be used.

To deduce the critical dimensions of a percolating network one can observe the survivability of a branch in the network. Define

$$F(x) = 1 - p_c + p_c G_1(x), \quad (12)$$

to be the degree distribution at criticality, i. e., the distribution of the number of occupied links leading to occupied nodes. Denote the number of nodes at a distance  $\ell$  along a branch by  $N_\ell$ . The distribution of such values can be fitted with a generating function  $N_\ell(x)$ . The generating functions for different layers satisfy

$$N_{\ell+1}(x) = F(N_\ell(x)). \quad (13)$$

At criticality the average number of nodes on the  $\ell$ th layer along a branch is 1, since a lower branching factor will lead to fast extinction of all branches, and a higher branching factor will give an infinite branch with a high probability. Therefore, the number of nodes at a distance  $\ell$  only along branches that survive at least  $\ell$  layers equals the average number of nodes in the  $\ell$ th layer, which is 1, divided by the fraction of branches surviving at least  $\ell$  layers. This number can be obtained by noticing that the probability of a branch to become extinct at the first  $\ell$  layers is given by  $N_\ell(0)$ , the probability of having 0 nodes in the  $\ell$ th layer. The average number of nodes at the  $\ell$ th layer of surviving branches is therefore  $m_\ell = 1/(1 - N_\ell(0))$ , and the fractal dimension can be obtained from the total number of nodes up to the  $\ell$ th layer in surviving branches,  $M_\ell = \sum_{i=1}^{\ell} m_i$ . The asymptotic behavior of  $N_\ell(0)$  can again be obtained and leads to  $M_\ell \propto \ell^{(\gamma-2)/(\gamma-3)}$  for  $3 < \gamma < 4$  and  $M_\ell \propto \ell^2$  for  $\gamma > 4$  and for ER networks. This implies

$$d_f = \begin{cases} \frac{\gamma-2}{\gamma-3}, & 3 < \gamma < 4, \\ 2, & \gamma > 4. \end{cases} \quad (14)$$

The size of the largest component at criticality can be obtained using the other critical exponents. As presented

above, the asymptotic behavior of the component size distribution is  $P(s) = s^{-\tau+1}$ , where  $P(s)$  is the probability of a node to belong to a component of size  $s$ . There are  $N$  nodes in the graph, and therefore it is expected that the size of the largest component of a graph,  $S$ , will be such that  $P(S) \sim N^{-1}$ . This leads to  $S \sim N^{1/(\tau-1)}$  and to

$$S \propto \begin{cases} N^{(\gamma-2)/(\gamma-1)}, & 3 < \gamma < 4, \\ N^{2/3}, & \gamma > 4. \end{cases} \quad (15)$$

### Optimal Paths and Minimum Spanning Trees

Communication in a network usually follows the shortest path from source to destination. The network structure and function usually depends only weakly on the space the network exists in. Hence, the path length is usually defined by the intrinsic properties of the network. The simplest definition of the path length between two nodes,  $a$  and  $b$ , in a network is the hop distance, i. e., the minimum number of links that need to be traversed in order to arrive from  $a$  to  $b$ . The average path length in networks has been studied extensively, and is known to be logarithmic in the size of the network for ER [5] and scale free graphs with  $\gamma > 3$  [29] and of order  $\log \log N$  in scale free graphs with  $\gamma < 3$  [15].

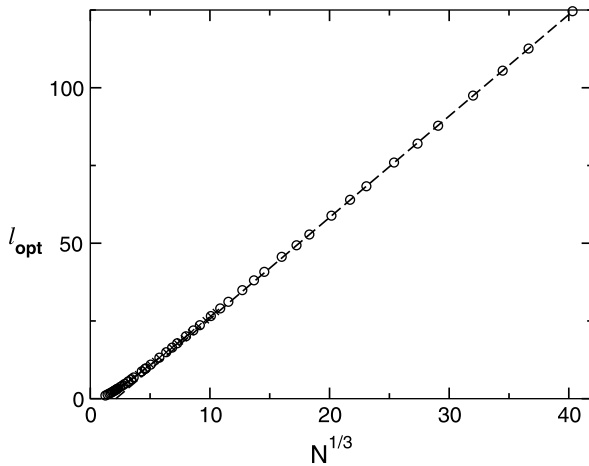
Another definition for the distance in a network can be given in case each link has some intrinsic length, or some intrinsic property (similar to energy in physics), termed “weight”, measuring the cost of using it. When the link length distribution is narrow (the “weak disorder” limit) the behavior of the optimal (lowest cost) path is expected to be very similar to that of the shortest path. However, when the distribution of link costs becomes very wide (the “strong disorder” limit) the behavior of the optimal path becomes very different from that of the shortest path [12,38].

The limit of strong disorder is observed clearly when the weight distribution is so wide, that the weight of each link is expected to be at least twice as large as the next highest link weight. This implies also that the weight of each link is larger than that of *all* links with lower weight. In this case paths can be compared by sorting the link weights along each path, and then comparing the lists of link weights by lexicographic order. It should be noted that the graph of all shortest paths also called “optimal paths” is a tree, i. e., has no loops. It is similar to the minimum spanning tree (MST), which presents the same behavior regardless of the weight distribution. Another similar case is high bandwidth information transmission in communication networks, where it may be desirable to transmit

through the path having the highest *minimal* bandwidth, to avoid bottlenecks.

An alternative method for reaching the optimal path tree is the bombing method. In this method the links in the network are removed one by one by order of decreasing weight. It is clear that removing a high weight link will not change any of the optimal paths, unless it makes the network disconnected, in which case it is not removed. Since the order of the link weights is random, so is the order of removal, and therefore this model is identical to random percolation, with the difference of refraining from removing links that make the network disconnected. Since at criticality percolation disintegrates the network into a collection of trees (or almost trees), the critical percolation component (cluster) is a subgraph of the optimal path tree.

This mapping of the optimal path and minimum spanning tree to a (restricted) percolation problem is very useful, since it allows the determination of the properties of these objects, based upon their similarity to the network at the critical point. From the results above in Sect. “Critical Exponents and Fractal Dimensions” the size of the largest component in ER networks is  $S \propto N^{2/3}$ . The chemical dimension is  $d_l = 2$ . Therefore,  $S \propto \ell^2$ , leading to the average hop distance between nodes on the critical components being  $\ell \propto N^{1/3}$ . Similarly for scale free networks with  $3 < \gamma < 4$ ,  $\ell \propto N^{(\gamma-3)/(\gamma-1)}$ . This presents a lower bound on the length of the optimal path in ER and scale free networks [7]. In fact, it is observed that the critical components are connected in a compact way i. e., through a small number of components in the optimal path tree (or MST) and this lower bound is actually exact. See Fig. 5.



Percolation in Complex Networks, Figure 5

The optimal distance  $l_{\text{opt}}$  as a function of  $N^{1/3}$  for ER graphs with strong disorder. After [7]

## Fragmentation of Social Networks

One of the most interesting questions in sociological networks is quantifying the collapse process of a network. Under certain circumstances it may happen that a network of friendship or acquaintance is fragmented into several components. It is desirable in many cases to quantify the fragmentation using a measure that is sensitive to the different possible partitions into fragments. Such a measure, developed in [6] can distinguish between different partitions of a network, based on the sizes of all components. This is especially important in small and medium size networks, and in non random fragmentation processes, where the fragments may contain several similarly sized components.

Notice that the phenomenon of network fragmentation due to link removal (acquaintance separation) or node removal (individuals leaving the social network) is similar to bond and site percolation, respectively. One of the main shortcomings of the percolation description in this case is the focus of percolation theory on random processes and in the limit of large system size (the “thermodynamic limit”). Here we will discuss the proposed fragmentation measure and its relation to the standard percolation measures.

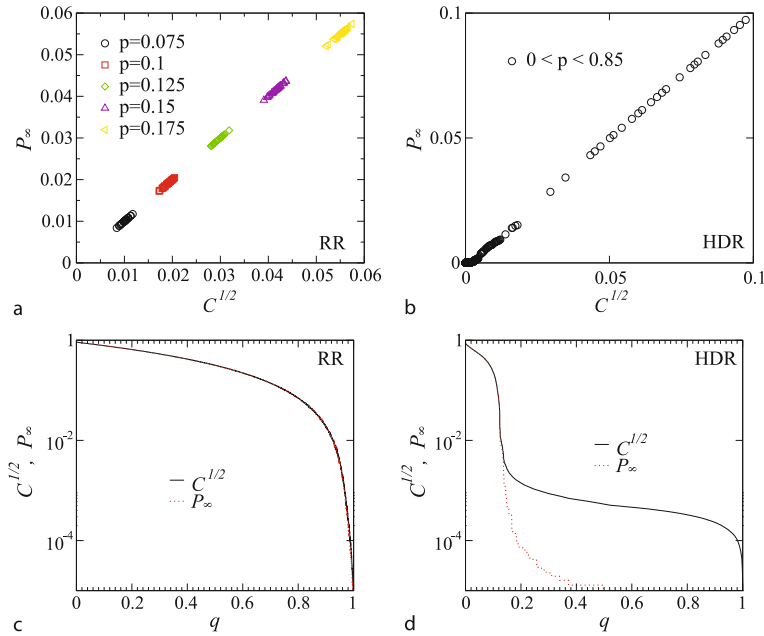
The suggested measure of fragmentation,  $F$ , is the number of pairs of nodes not reachable from each other, divided by the total number of pairs in the network [6]. Since nodes that are reachable from each other belong to the same component, this is equivalent to the following definition:

$$F = 1 - \frac{\sum_{i=1}^n s_i(s_i - 1)}{N(N - 1)} \equiv 1 - C, \quad (16)$$

where  $s_i$  is the size of the  $i$ th component and  $n$  is the number of components in the graph.

In [10] the relation between percolation theory and this fragmentation measure was studied. As discussed above, in the limit of  $N \rightarrow \infty$  above the percolation threshold there is a large gap between the size of the giant component, which is of order  $N$  and the size of the second largest component, which is usually of logarithmic size. At the threshold, the size of the largest component is of order  $N^{2/3}$  for ER networks and some power of  $N$  for scale free networks, and the component size distribution is continuous. Below the percolation threshold the distribution is continuous again. However, the size of the largest component is even smaller (logarithmic in ER networks). Equation (16) therefore can be presented in the following equivalent form in the limit of large  $N$

$$F \approx 1 - P_\infty^2 - \frac{\sum_s n(s)s(s-1)}{N^2} \approx 1 - P_\infty^2 - \frac{\langle s \rangle}{N}. \quad (17)$$



**Percolation in Complex Networks, Figure 6**

Comparison of  $P_\infty$  and  $C \equiv 1 - F$  for random removal (a and c) and high degree removal (b and d) of links in a real social network of working relations in Sweden. After [10]

Notice that  $\langle s \rangle$  represents the average component sampled over all nodes, rather than over all clusters. This gives a larger weight to larger components. In fact, since  $\tau = 2.5$  in ER networks, and  $\tau < 3$  for all  $\gamma > 3$  and that  $P(s)$ , the probability of a node to belong to a component of size  $s$  scales as  $P(s) \sim s^{-\tau+1}$ , it follows that  $\langle s \rangle$  diverges similarly to  $S$  at the threshold.

For finite networks in particular  $F$  has the advantage that it presents some measure of the fragmentation process both above and below the critical threshold. A comparison of  $P_\infty$  and  $F$  as measures of fragmentation can be seen in Fig. 6.

### Future Directions

Different types of percolation can be defined and studied on random networks. Random percolation has special features in terms of the threshold and the critical exponents in scale free networks. The percolation theory of networks has many applications in epidemiology, network robustness, social networks analysis, and communication network efficiency.

Several open questions still remain regarding percolation theory in networks and its applications. For scale free networks with  $2 \leq \gamma \leq 3$  it seems that a phase transition still exists with a threshold that approaches zero as a function of  $N$ . Although some progress has been made in this

direction (see, e.g., [25]), the nature of this transition is not completely well understood and understanding it may enable the understanding of properties such as the optimal path behavior in such networks (which, from numerical results, seems to behave logarithmically).

Other important topics, which are not yet fully understood, is the question of optimal network design, and optimized attack strategies. In optimal network design, an attempt is made to find the parameters (such as degree distribution, correlations, or clustering) that produce a network class, with optimal percolation properties, including minimum random or targeted percolation thresholds, good near critical behavior, etc. Optimized attack strategies attempt to bring the network to the percolation threshold (or any other desired point) with the minimum number of removed nodes or links. Targeted attacks are usually very efficient in achieving percolation. However, more efficient methods can be conceived.

### Bibliography

#### Primary Literature

1. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
2. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382

3. Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
4. Bollobás B (1980) A probabilistic proof of an asymptotic formula for the number of labeled regular graphs. *Eur J Comb* 1:311–316
5. Bollobás B (1985) *Random graphs*. Academic Press, London
6. Borgatti SP (2006) Identifying sets of key players in a network. *Comput Math Organ Theor* 12:21–34
7. Braunstein LA, Buldyrev SV, Cohen R, Havlin S, Stanley HE (2003) Optimal paths in disordered complex networks. *Phys Rev Lett* 91:168701
8. Bunde A, Havlin S (1996) *Fractals and disordered system*. Springer, New York
9. Callaway DS, Newman MEJ, Strogatz SH, Watts DJ (2000) Network robustness and fragility: percolation on random graphs. *Phys Rev Lett* 85:5468–5471
10. Chen Y, Paul G, Cohen R, Havlin S, Borgatti SP, Liljeros F, Stanley HE (2007) Percolation theory applied to measures of fragmentation in social networks. *Phys Rev E* 75:046107
11. Cieplak M, Maritan A, Banavar JR (1999) Optimal paths and growth process. *Physica A* 266:291–298
12. Cohen R, ben-Avraham D, Havlin S (2003) Percolation Critical Exponents in Scale Free Networks. *Phys Rev E* 66:036113
13. Cohen R, Erez K, ben-Avraham D, Havlin S (2000) Resilience of the Internet to Random Breakdown. *Phys Rev Lett* 85:4626–4628
14. Cohen R, Erez K, ben-Avraham D, Havlin S (2001) Breakdown of the Internet under Intentional Attack. *Phys Rev Lett* 86:3682–3685
15. Cohen R, Havlin S (2003) Scale free networks are ultrasmall. *Phys Rev Lett* 90:058701
16. Cohen R, Havlin S, ben-Avraham D (2003) Efficient Immunization Strategies for Computer Networks and Populations. *Phys Rev Lett* 91:247901
17. Dezső A, Barabási AL (2002) Halting viruses in scale free networks. *Phys Rev E* 65:055103
18. Dorogovtsev SN, Mendes JFF (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford
19. Erdős P, Rényi A (1959) On random graphs. *Publ Math* 6:290–297
20. Erdős P, Rényi A (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* 5:17–61
21. Erdős P, Rényi A (1961) On the strength of connectedness of a random graph. *Acta Math Sci Hung* 12:261–267
22. Grassberger P (1983) On the critical behavior of the general epidemic process and dynamical percolation. *Math Biosci* 63:157–172
23. Krapivsky PL, Redner S (2002) A Statistical Physics Perspective on Web Growth. *Comput Netw* 39:261–276
24. Havlin S, Nossal R, (1984) Topological properties of percolation cluster. *J Phys A* 17:L427–L432
25. Lee DS, Goh KI, Kahng B, Kim D (2004) Evolution of scale-free random graphs: Potts model formulation. *Nucl Phys B* 696:351–380
26. Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. *Random Struct Algorithms* 6:161–179
27. Newman MEJ (2002) The spread of epidemic disease on networks. *Phys Rev E* 66:016128
28. Newman MEJ (2003) Structure and function of complex networks. *SIAM Rev* 45:167–256
29. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Phys Rev E* 64:026118
30. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale free networks. *Phys Rev Lett* 86:3200–3203
31. Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. *Phys Rev E* 63:066117
32. Pastor-Satorras R, Vespignani A (2003) *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, Cambridge
33. Rapoport A (1957) A contribution to the theory of random and biased nets. *Bull Math Biophys* 19:257–271
34. Sander LM, Warren CP, Sokolov IM, Simon C, Koopman J (2002) Percolation on heterogeneous networks as a model for epidemics. *Math Biosci* 180:293–305
35. Stauffer D, Aharony A (1991) *Introduction to percolation theory*. Taylor and Francis, London
36. Schwartz N, Porto M, Havlin S, Bunde A (1999) Optimal path in weak and strong disorder. *Physica A* 266:317–321
37. Watts DJ, Strogatz SH (1998) Collective dynamics of “small world” networks. *Nature* 393:440–442
38. Wilf HS (1994) *Generatingfunctionology*. Academic Press, San Diego

## Books and Reviews

- Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
- Barabási AL (2002) *Linked: The New Science of Networks*. Perseus, Cambridge
- Bornholdt S, Schuster HG (2002) *Handbook of Graphs and Networks*. Wiley-VCH, Berlin
- Cohen R, Havlin S (2008) *Complex Networks: Structure, Stability and Function*. Cambridge University Press, Cambridge (in press)
- Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. *Adv Phys* 51:1079–1187
- Dorogovtsev SN, Mendes JFF (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford
- Newman MEJ (2003) Structure and function of complex networks. *SIAM Rev* 45:167–256
- Newman MEJ, Barabási AL, Watts DJ (2006) *The Structure and Dynamics of Networks*. Princeton University Press, Princeton
- Pastor-Satorras R, Vespignani A (2003) *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, Cambridge

## Percolation, and Faults and Fractures in Rock

PIERRE M. ADLER<sup>1</sup>, JEAN-FRANÇOIS THOVERT<sup>2</sup>,  
VALERI V. MOURZENKO<sup>2</sup>

<sup>1</sup> UPMC-Sisyphe, Paris, France

<sup>2</sup> CNRS-LCD, Chasseneuil du Poitou, France



## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Fracture Networks  
 Percolation of Fracture Networks  
 Determination of the Dimensionless Density  
   from Experimental Data  
 Role of the Dimensionless Density  
   in Other Geometrical Properties and Permeability  
 Future Directions  
 Bibliography

## Glossary

**Dimensionless density** The dimensionless density is the number of objects per excluded volume.

**Excluded volume** The *excluded volume*  $V_{\text{ex}}$  of an object is defined as the volume surrounding it, in which the center of another object must be in order for them to intersect.

**Fracture network** A fracture network is generally defined as a set of individual fractures which may or may not intersect.

**Percolation and percolation threshold** Percolation is defined as the existence of a spanning connected cluster in the fracture network. Percolation occurs when the number of fractures per unit volume is equal or larger than a certain value called the percolation threshold.

**Plane convex fractures** A plane fracture is convex if for any points  $A$  and  $B$  which belong to the fracture, all the points of the segment  $AB$  belong to the fracture.

## Definition of the Subject

The study of fractured porous media is of great practical and theoretical importance. It has been first generated by the fact that the presence of fractures can change completely the macroscopic properties of porous media which are present for instance in oil reservoirs, aquifers or waste repositories. The first contributions to this subject were made from very different standpoints by Barenblatt and coworkers [7], Conrad and Jacquin [13], and Witherspoon and coworkers (see for instance [22]).

In the eighties, these studies were renewed by concepts such as percolation and fractals. Fracture networks were first addressed in the framework of continuum percolation by [12] and [4].

Since the mid nineties, important progress have been made in this field thanks to systematic numerical experi-

ments which can be rationalized by using the concept of excluded volume.

## Introduction

Knowledge of geometrical properties of fracture networks is crucial to the understanding of flow and other transport processes in geological formations, both at small and large scales; introduction of fractures in a porous rock matrix seriously alters the macroscopic properties of the formation. Moreover, studies of fracture geometries during the last 30 years show that naturally occurring geological fractures exist on scales ranging from a few mm to several kilometers [33]. Therefore, fracture networks are likely to influence the transports on a large range of scales. Because of their importance, fracture networks are studied and applied in various areas such as oil and gas recovery, hydrology, nuclear waste storage and geothermal energy exploitation.

Geological fractures can be defined as discrete discontinuities within a rock mass; these breaks are characterized by the fact that their local aperture (defined as the local distance between the two surfaces which limit the fracture) is significantly smaller than their lateral extent; in other words, when they are viewed from far away, fractures can be assimilated to surfaces of discontinuity; in most cases, these surfaces are relatively plane. Fractures have varying degrees of aperture, and may in some cases be completely closed either because of deposition of material induced by fluid flow, or by displacements of the matrix.

An important property of these fracture sets or fracture network is their connectivity and their percolation properties. If a network percolates, fluid can circulate only through it and most likely much more rapidly than in the surrounding porous medium itself. Connectivity studies of fracture networks were initiated in 3D by Charlaix et al. [12] and Balberg [4].

The purpose of this paper is to provide a complete and updated view of the percolation properties of fracture networks in rocks. It is organized as follows. In Sect. “[Fracture Networks](#)”, fractures are modeled as plane convex polygons which enables the introduction of the concept of excluded volume  $V_{\text{ex}}$ . This volume is a simple function of the surface and perimeter of the fractures, and it enables to introduce a dimensionless fracture density  $\rho'$  which is defined as the number of fractures per excluded volume. The tools necessary for the numerical study of the percolation thresholds are detailed and applied to mono- and poly- disperse fracture networks. It is shown that when expressed in terms of  $\rho'$ , the percolation threshold does not

depend anymore on the fracture shapes. This crucial property is presented and discussed.

Section “[Determination of the Dimensionless Density from Experimental Data](#)” is devoted to the determination of the dimensionless density from experimental data. In most cases, these data are based on 1D and 2D measurements of fracture traces along boreholes or on exposed outcrops. These measurements necessitate extrapolation by stereological techniques to three dimensions. Significant progress can be made for plane convex fractures. Some recent applications of the methodology are given.

Finally, the independence of the dimensionless percolation threshold on the fracture shape can be extended to other properties such as other geometric properties and the macroscopic permeability of fractured rocks. These extensions are summarized in Sect. “[Role of the Dimensionless Density in Other Geometrical Properties and Permeability](#)”.

## Fracture Networks

A fracture network is generally defined as a set of individual fractures which may or may not intersect.

On a scale large with respect to the fracture aperture, fractures are usually modeled as convex, finite polygons possibly based on an embedding disk as shown in Fig. 1a. This is only a simplifying assumption which however provides a standard starting point for studying fracture networks. Convex polygons can be used to analyze shape and area dependencies of geometrical and topological features in the fracture systems in a systematic way.

The individual fractures are characterized by their orientation. This orientation is usually given by two unit vectors  $\mathbf{n}$  and  $\mathbf{m}$  (cf. Fig. 1).  $\mathbf{n}$  is the normal to the fracture

plane;  $\mathbf{m}$  gives the orientation of the polygon in the fracture plane.

The simplest model consists of a network in which all fractures have the same shape and are inscribed in a circle with a given radius  $R$ . The normal vectors  $\mathbf{n}$  are uniformly distributed on the unit sphere. The density  $\rho$  of this isotropic monodisperse network is defined as the number of fractures per unit volume. An illustration of such a fracture system is shown in Fig. 2a.

Next consider three-dimensional networks made up of polydisperse fractures with plane polygonal shapes. These polygons may be regular or not, but all their vertices are supposed to lie on their circumscribed circle, whose radius  $R$  provides a measure of their size. In agreement with many observations of fractured rocks [2], the statistical distribution of the fracture sizes is supposed to be given by a power law

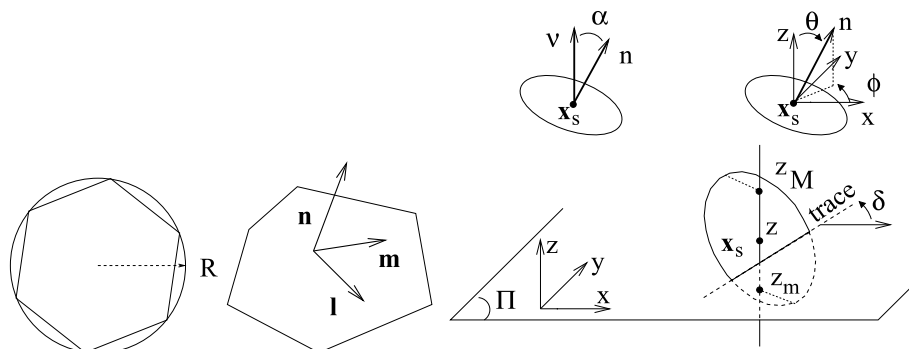
$$n(R) = \alpha R^{-a} \quad (1)$$

where  $n(R)dR$  is the probability of fracture radii in the range  $[R, R + dR]$ ;  $\alpha$  is a normalization coefficient, and the exponent  $a$  ranges between 1 and 5. In practice,  $R$  may vary over a large interval which can span five orders of magnitude, from the size  $R_m$  of the microcracks to the size  $R_M$  of the largest fractures in the system. The normalization condition implies that  $\alpha$  verifies

$$\alpha = \frac{a-1}{R_m^{1-a} - R_M^{1-a}} \quad (a \neq 1); \quad (2)$$

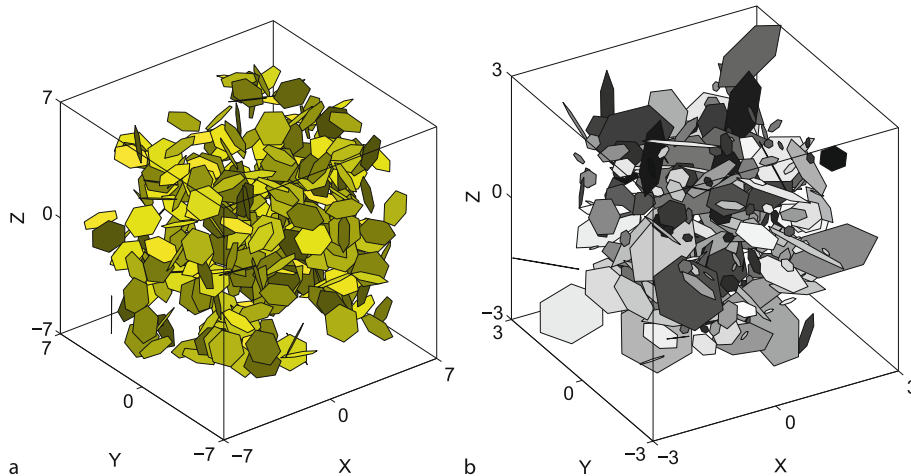
$$\alpha = \frac{1}{\ln R_M - \ln R_m} \quad (a = 1).$$

The definition of the network density  $\rho$  for polydisperse networks should be modified. To this end, we introduce the volumetric number density of fracture per frac-



**Percolation, and Faults and Fractures in Rock, Figure 1**

**Notations.** Convex polygons such as hexagons are created within a circle of radius  $R$  (a). This polygon requires two unit vectors  $\mathbf{n}$  and  $\mathbf{m}$  to be oriented in space (b);  $\mathbf{l}$  is a unit vector perpendicular to  $\mathbf{n}$  and  $\mathbf{m}$ . c illustrates the notations which are mostly used in Sect. “[Determination of the Dimensionless Density from Experimental Data](#)”



**Percolation, and Faults and Fractures in Rock, Figure 2**

Examples of three-dimensional fracture networks. **a** Monodisperse network made of identical polygons. The volume of size  $L^3$  contains 495 hexagons;  $L = 12R$  where  $R$  is the radius of the circle in which the hexagon is inscribed. **b** Polydisperse network of hexagonal fractures, with  $L' = 4$ ,  $a = 1.5$ ,  $R'_m = 0.1$ , which contains  $N_{fr} = 300$  fractures ( $\rho'_{21} = 1.25$ ,  $\rho'_3 = 2.44$ ). The unit for the coordinates is  $R_M$

ture size  $F(R)$ ,

$$F(R) = \rho n(R) \quad (3)$$

where  $F(R)dR$  is the number of fractures with radius in the range  $[R, R + dR]$  per unit volume.

An example of such polydisperse networks is given in Fig. 2b.

## Percolation of Fracture Networks

### General Considerations on Continuum Percolation

**Continuum Percolation** Percolation, i. e., the existence of a spanning connected cluster in the fracture network, is a crucial topological property which conditions many other geometrical or transport properties of the network.

Percolation of discrete sites or bonds lattices has been closely studied (see, e. g. [32,37]). In these lattices, the sites or bonds are occupied with a probability  $p$ , which can be interpreted as a concentration. In large systems, percolation occurs when  $p$  exceeds a critical value  $p_c$ , known as the percolation threshold, which depends on the underlying lattice structure. For  $p$  close to  $p_c$ , however, many geometrical or transport coefficients are known to scale as power laws of the difference  $p - p_c$ , according to the standard form

$$X \propto (p - p_c)^\alpha. \quad (4)$$

The quantity  $X$  may represent the correlation length, the fraction of sites connected to the infinite cluster, or the

conductivity of the system. Different exponents are associated with the various quantities, but each is generally believed to be universal, i. e., insensitive to the details of the underlying lattice.

It is, of course, tempting to try to transpose this theoretical framework to the problem at hand. It is intuitively obvious that a fracture network will start percolating if some critical concentration is reached. The main difficulty is to define an equivalent of the probability  $p$  in discrete lattices. As shown below, this can be done by using the concept of excluded volume, introduced by [6] in the context of fracture networks.

Fracture networks belong to the general class of continuum percolation systems. Applications of continuum percolation concepts to geophysical problems have been reviewed by [9]. Continuum percolation differs from lattice percolation in several respects. First, the occupancy probability  $p$  in a discrete lattice ranges between 0 and 1, which means that there is a maximal concentration; the filling of the system can be defined relative to this upper bound. In continuum percolation, there is generally no such upper bound. For instance, there is ideally no upper limit to the degree of fracturation of a piece of rock. Consequently, the relative concentration  $p$  has to be replaced by a volumetric density. Second, any site or bond in lattice percolation cannot have more than a maximum number  $z$  of neighbors, called the lattice coordination number, whereas there is no limitation to the number of intersections for a fracture in a network. Other differences result from the variable lengths and orientations of the

bonds, in contrast with the discrete set of values imposed by a lattice, which may be significant for transport properties (see [5]).

Note that in this section we only consider “large” systems, i. e., the size of the objects in the percolation system may have a broad distribution, but the overall domain extension is supposed to widely exceed the size of the largest objects it contains. This condition may sometimes be difficult to fulfil; natural fracture networks often involve large-scale faults, which may in themselves ensure percolation if they cross the domain of investigation. [11] and later [24] considered such broad size-distributions where the probability of a spanning single fracture is non-zero.

In view of the previous considerations, two definitions of the system concentration appear possible. One is volumetric, quantified by the average number of objects in a reference volume; the other is topological, defined as the average number of connections with surrounding objects. These two definitions are nicely reconciled by the introduction of the concept of excluded volume.

The *excluded volume*  $V_{\text{ex}}$  of an object was defined by [6] as the volume surrounding it, in which the center of another object must be in order for them to intersect. We first discuss the simplest case of populations of identical objects, with volume  $V$ . For example, the excluded volumes of a sphere with volume  $V$  in 3D and of disks with area  $A$  in 2D are

$$\begin{aligned} V_{\text{ex}} &= 8V \quad \text{for spheres;} \\ A_{\text{ex}} &= 4A \quad \text{for disks in the plane.} \end{aligned} \quad (5)$$

These equations are also valid for any object with convex shape, if all the objects in the population have identical orientations.

If the objects are anisotropic and have distributed orientations, the excluded volume has to be averaged over all possible relative orientations of the intersecting objects.

Now suppose that the volumetric density of objects per unit volume is  $\rho$ . It is natural to use  $V_{\text{ex}}$  as a reference volume, and we may define the dimensionless density  $\rho'$  as the number of objects per volume  $V_{\text{ex}}$

$$\rho' = \rho V_{\text{ex}}. \quad (6)$$

On the other hand, the definition of  $V_{\text{ex}}$  implies that  $\rho'$  is also the average number of intersections per object, if they are randomly located according to a Poisson process. Therefore, given the shape of the object and its orientation distribution (and thus  $V_{\text{ex}}$ ), the definition (6) incorporates both the volumetric and topologic aspects mentioned above.

It should be emphasized however, that the definition of the excluded volume is meaningful only if the object locations are uniformly distributed in space. If there are spatial correlations, they should be replaced by a spatial integral of the pair separation distribution function (see for instance [14] for applications to the physics of liquids).

### Calculation of the Excluded Volume for Plane Convex Fractures

A general expression for the excluded volume was established very early in the context of statistical mechanics by [19], for isotropically oriented objects. For two three-dimensional convex objects  $A$  and  $B$  with volumes  $V_A$  and  $V_B$ , areas  $A_A$  and  $A_B$  and surface averaged mean radius of curvature  $R_A$  and  $R_B$ , [19] obtained the mutual exclusion volume

$$V_{\text{ex},AB} = V_A + V_B + (A_A R_B + A_B R_A). \quad (7)$$

This expression can then be averaged over the distributions of object shapes and sizes. For equal spheres, Eq. (5) is obtained. For flat convex objects randomly oriented in space with perimeters  $P_A$  and  $P_B$ , it is reduced to [12]

$$V_{\text{ex},AB} = \frac{1}{4} (A_A P_B + A_B P_A). \quad (8)$$

On averaging (8) over the size distribution of objects with identical shapes, one obtains

$$V_{\text{ex}} = \frac{1}{2} \langle A \rangle \langle P \rangle \quad (9)$$

where  $\langle \cdot \rangle$  is the statistical average. If  $A$  and  $B$  are identical, (9) yields

$$V_{\text{ex}}^{\text{iso}} = \frac{1}{2} A P. \quad (10)$$

If the population of polygons is not isotropic and has a probability distribution  $n(f)$ , which may involve the shape or the size of the polygons, the average of (8) yields

$$\begin{aligned} V_{\text{ex}}^{\text{iso}} &= \frac{1}{4} \iint n(F_1) n(F_2) (A_1 P_2 + A_2 P_1) dF_1 dF_2 \\ &= \frac{1}{2} \langle A \rangle \langle P \rangle \end{aligned} \quad (11)$$

where  $\langle A \rangle$  and  $\langle P \rangle$  are the average area and perimeter. Alternatively, the polygon orientation may be incorporated into  $n(f)$  and a general expression of  $V_{\text{ex}}$  can be obtained.

### Determinations of Continuum Percolation Thresholds

The percolation thresholds of various simple continuous systems have been determined, since the pioneering papers of [35] and [27]. These early works were reviewed



**Percolation, and Faults and Fractures in Rock, Table 1****Thresholds  $\rho'_c$  for various continuum percolation systems in  $d$  dimensions**

$d$	Object type	$\rho'_c$	$d$	Object type	$\rho'_c$
2	Orthogonal sticks	3.2	3	Orthogonal elongated rods	0.7
2	Randomly oriented sticks	3.6	3	Randomly oriented elongated rods	1.4
2	Disks or parallel objects	4.5	3	Orthogonal squares	2.0
			3	Randomly oriented squares	2.46
			3	Spheres or parallel objects	2.80

by [6] and [5]. A few examples are given in Table 1, for monodisperse objects in a  $d$ -dimensional space. The critical concentration is described in terms of the average number  $\rho'_c$  of connections per object.

The influence of the orientation distribution was investigated by [29,30], and [6]. For sticks with constant length in the plane, [29] has shown that  $\rho'_c$  is identical for uniform orientation distributions in any angular sector and equal to the value 3.6 for an isotropic distribution. On the other hand, the value 3.2 for orthogonal sticks is also valid for any bimodal orientation distribution. By considering three-dimensional systems [6], also conclude that the total excluded volume at percolation is independent of the degree of anisotropy. [4] proposed a set of bounds which correspond to orthogonal and parallel object systems

$$3.2 \leq \rho'_c \leq 4.5 \quad d = 2; \quad 0.7 \leq \rho'_c \leq 2.8 \quad d = 3. \quad (12)$$

All these results were obtained by numerical simulations. One should also mention the heuristic criterion developed by [3]. They define the average “bonding distance”  $l$  as the mean distance between connected objects, which is essentially the gyration radius of the excluded volume

$$l^2 = \frac{1}{V_{\text{ex}}} \int_{V_{\text{ex}}} r^2 d^3 \mathbf{r}. \quad (13)$$

Note that  $l$  does not depend on the density of objects. They then postulate that percolation occurs when the average distance  $L_d$  between objects with at least two neighbors is smaller than or equal to  $2l$ . To evaluate  $L_d$ , they note that the number  $k$  of connections to a given object corresponds to a Poisson distribution

$$\text{Pr}(k) = \frac{\rho'^k}{k!} e^{-\rho'}. \quad (14)$$

Therefore, the density  $\rho_2$  of objects with at least two neighbors is

$$\rho_2 = \rho \left[ 1 - (1 + \rho') e^{-\rho'} \right]. \quad (15)$$

Thus, an estimate of  $L_d$  follows from

$$\frac{4}{3} \pi \left( \frac{L_d}{2} \right)^3 = \frac{1}{\rho_2}. \quad (16)$$

An equation for the critical concentration  $\rho'_c$  can be directly deduced from the statement that  $L_d = 2l$ . Although the argument is not substantiated, it is quite successful. It yields directly  $\rho'_c = 2.80$  for spheres. In two dimensions,  $L_d$  is replaced by the average distance between objects with at least 5 neighbors.

An interesting feature of this argument is that it can be easily generalized to account for spatial correlations. If  $\rho g(r)$  denotes the probability density of finding an object center at a distance  $r$  from an object located at the origin, the bonding distance is defined by the weighted average

$$l^2 = \frac{\int_{V_{\text{ex}}} r^2 g(r) d^3 \mathbf{r}}{\int_{V_{\text{ex}}} g(r) d^3 \mathbf{r}}. \quad (17)$$

Similarly, the average number of bonds per object appears as

$$\rho' = \rho \int_{V_{\text{ex}}} g(r) d^3 \mathbf{r}. \quad (18)$$

Using these two definitions, an equation for  $\rho'_c$  can be obtained. Its predictions were successfully compared by [3] to numerical simulations for systems of hard-core spheres with or without interaction potentials.

Only monodisperse objects have been addressed so far in this subsection. For polydisperse populations, there seems to be some confusion in the literature. For flat objects, the statistical derivation of the excluded volume in Sect. “Calculation of the Excluded Volume for Plane Convex Fractures” quite naturally yielded the averages (9) or (11), which account for the sizes of the two intersecting objects. For isotropic populations of segments with length  $l$  in the plane or disks with radius  $R$  in space, for instance, the averages can be expressed as

$$V_{\text{ex}} = \frac{2}{\pi} \langle l \rangle^2 \text{ segments}, \quad d = 2 \quad (19a)$$

$$V_{\text{ex}} = \frac{\pi^2}{8} \langle R^2 \rangle \langle R \rangle \text{ disks}, \quad d = 3. \quad (19b)$$

However, another type of average has been proposed by [6], namely,

$$W_{\text{ex}} = \frac{2}{\pi} \langle l^2 \rangle \text{ segments}, \quad d = 2 \quad (20a)$$

$$W_{\text{ex}} = \frac{\pi^2}{8} \langle R^3 \rangle \text{ disks, } d = 3. \quad (20b)$$

On the basis of the numerical simulations of [29,30], [6] and others claim that the average bond number for polydisperse objects is not given by Eq. (6) but instead by

$$\rho'' = \rho W_{\text{ex}}. \quad (21)$$

However, a careful examination of [29]'s data shows that they correspond very accurately to Eq. (6) with (19a). Finally, the derivations of [8] are based on (19b) and yield consistent results.

Actually [29], and [30] showed that  $\rho'_c$  is not invariant for similar systems of segments in the plane with various degrees of polydispersity, while  $\rho''_c$  is. [11] also observed that  $\rho''_c$  is roughly constant for very broad power-law segment size distributions. The profound meaning of this observation is that continuum percolation is not determined only by the average coordination, when connections over various ranges may coexist. As suggested by [28], this is probably because contacts between objects too close to each other are redundant to percolation.

To summarize, the density  $\rho'$  based on  $V_{\text{ex}}$  resulting from the averages (9), (11), or (19a) is always equal to the mean number of intersections per object, but it cannot be used to relate the percolation thresholds of mono- and polydisperse systems. The alternative definition  $\rho''$  in (21) is very successful in this respect and is going to be generalized as  $\rho'_3$  in (29).

## Methods

The three main tools necessary for the numerical study of the percolation properties of the fracture network models are summarized in this section.

First, the medium is assumed to be spatially periodic on a large scale. A detailed description of spatially periodic media is given by [1], and only the main characteristics of these models are briefly repeated here. The geometrical and physical properties of the system under investigation are invariant under the translations

$$\mathbf{R}_i = i_1 \mathbf{l}_1 + i_2 \mathbf{l}_2 + i_3 \mathbf{l}_3 \quad (22)$$

where  $\mathbf{i} = (i_1, i_2, i_3) \in \mathbb{Z}^3$ , and where  $\mathbf{l}_1$ ,  $\mathbf{l}_2$  and  $\mathbf{l}_3$  define a unit cell where the system is studied. The entire space is tiled by replicas of this unit cell, translated by  $\mathbf{R}_i$ . All the studies presented in this chapter are performed in cubic unit cells where  $|\mathbf{l}_1| = |\mathbf{l}_2| = |\mathbf{l}_3| = L$ .

Spatial periodicity implies that fractures may cross the imaginary unit cell boundaries, and reach the neighboring cells of the periodic medium. Therefore, for polydisperse

fractures,  $R_M$  should be at least smaller than  $L/2$ . Moreover, in order to represent a homogeneous medium by a periodic model, one has to set the unit cell size much larger than any finite characteristic length scale in the system. Practically speaking, because of the finite size effects which will be discussed in Sect. "Methods",  $R_M$  is at least smaller than  $L/4$ .

Second, the networks are characterized by a graph which provides all the necessary relations and information. This graph, denoted by  $\Gamma_1$ , consists of vertices which correspond to the polygons, and edges which correspond to the intersection between polygons.  $\Gamma_1$  will be used to study network percolation as a function of fracture shape, distribution and density, as well as to characterize the topological features of the percolating components of the networks.

The information relative to the intersections is stored in the graph  $\Gamma_1$ . Since the networks considered here are spatially periodic, intersections of a polygon  $P_1$  with the periodic replicas of a polygon  $P_2$  in the 26 neighboring unit cells have to be checked as well. Once the intersections have been identified, the edges of  $\Gamma_1$  are known, and  $\Gamma_1$  can be set up.

Third, in order to estimate the percolation thresholds, the classical finite-size scaling method described in [37] is used. The percolating system is studied for various cell sizes  $L$ . For given values of  $L$  and  $\rho$ , the probability  $\Pi_L(\rho)$  of having a percolating cluster is derived from numerous realizations of the system. Then, the numerical data are used to estimate  $\rho_{Lc}$  (the value for which  $\Pi_L(\rho) = 1/2$ ) and an estimate of the width  $\Delta_L$  of the transition region of  $\Pi_L(\rho)$ .  $\Pi_L(\rho)$  was fitted with an error function of the form

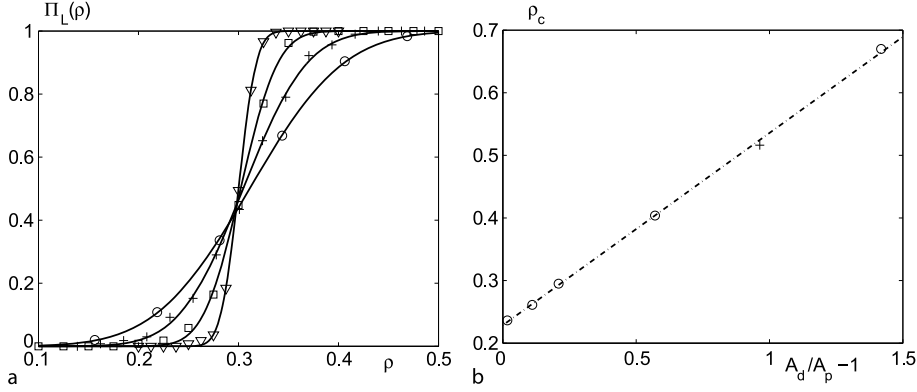
$$\Pi_L(\rho) = \frac{1}{\sqrt{2\pi}\Delta_L} \int_{-\infty}^{\rho} \exp \left\{ -\frac{(\xi - \rho_{Lc})^2}{2(\Delta_L)^2} \right\} d\xi \quad (23)$$

where  $\rho_{Lc}$  and  $\Delta_L$  are fit parameters. Once they have been evaluated for several values of  $L$ , the asymptotic value  $\rho_{Lc}$  for infinite systems  $\rho_c$  can be derived from the two scaling relations

$$\rho_{Lc} - \rho_c \propto L^{-1/\nu} \quad \Delta_L \propto L^{-1/\nu}. \quad (24)$$

## Monodisperse Fractures

This case was addressed by [18]. Since the computer time increases proportionally to the square of the number  $N$  of objects,  $L$  (measured in units of the disk radius  $R$ ) was kept below 16 in this early contribution. Despite the small cell sizes, the scaling laws (24) are well verified, which justifies



**Percolation, and Faults and Fractures in Rock, Figure 3**

**a** The probability of percolation  $\Pi_L(\rho)$  vs the density  $\rho$  of fractures in fracture networks created by equal sized, regular hexagons. Data are for sample sizes  $L/R = 4(\circ), 6(+), 10(\square), 20(\nabla)$ . The solid lines are the fitted error functions. **b** The percolation thresholds  $\rho_c$  for regular polygons ( $\circ$ ) and rectangles with  $a/b = 0.5(+)$  vs  $(A_d/A_p - 1)$ . The linear fit ( $-\cdot-\cdot-$ ) yields  $\rho_c = 0.231 \pm 0.002$  for disks

the extrapolations of  $\rho_c$  at  $L \rightarrow \infty$ . The polygons were created, and intersections identified. Percolation was checked in all possible directions  $x, y$  and  $z$ . Periodic boundary conditions were applied to the 3d graph during this search; this means that a cluster must touch two opposite faces of the unit cell, and in addition contain fractures intersecting one another across the faces.

An example of the plots of the estimated  $\Pi_L(\rho)$  data points is given in Fig. 3a, together with the fitted error functions. Plots of  $\ln(\Delta_L)$  vs  $\ln(L/R)$  were used to obtain the critical exponent  $\nu$ . The various polygons are expected to belong to the same universality class, and  $\nu$  was expected to be the same in all cases. Values were in the range  $\nu = 1.011 \pm 0.044$ . The plots of  $\rho_{Lc}$  vs  $\Delta_L$  were extrapolated for  $\Delta_L \rightarrow 0$  to find  $\rho_c$  and these extrapolations are shown in Fig. 3b as functions of the shape factor  $A_d/A_p - 1$  where  $A_d$  is the area  $\pi R^2$  of the circumscribed disk.

These results can be analyzed in terms of the average number of intersections per fracture  $\rho'$ . (10) can be applied to networks made of identical polygons

$$\frac{V_{\text{ex}}}{R^3} = \pi^2 \left( \frac{N_v}{\pi} \right)^2 \cos \left( \frac{\pi}{N_v} \right) \sin^2 \left( \frac{\pi}{N_v} \right), \quad (\text{regular } N_v\text{-polygons}) \quad (25a)$$

$$\frac{V_{\text{ex}}}{R^3} = \frac{8a(a+1)}{(a^2+1)^{3/2}}, \quad (\text{rectangles with aspect ratio } a) \quad (25b)$$

The resulting values of  $\rho'_c$  are remarkably constant (cf. [18]). For all the fracture networks, including the cases

of anisotropic (rectangular) polygons,  $\rho'_c$  is within the range

$$\rho'_c = 2.26 \pm 0.04. \quad (26)$$

Note that (26) concords with the limits (12) set up by [4] for 3d systems.

To summarize, this set of numerical results suggests that the percolation threshold of a network of identical Poissonian polygons has a universal value, expressed as Eq. (26).

### Polydisperse Fractures

Since natural fracture networks are likely to have more complex size and shape distributions, the extension of (26) to these cases is of great interest. The key for this extension is the definition of a proper averaging procedure for the excluded volume.

The fracture size  $R$  is always supposed to follow the power law (1). Moreover, fractures of various shapes  $S$  are considered as well as mixtures of shapes. The three length scales  $R_m, R_M$  and  $L$  define two dimensionless ratios

$$R'_m = \frac{R_m}{R_M}, \quad L' = \frac{L}{R_M}. \quad (27)$$

Moreover, it will be shown below that global connectivity (percolation) is no longer controlled solely by the local one (mean coordination), in the case of size polydispersity, and the definition of the percolation parameter has to be generalized. Since shape effects are well accounted for by  $\langle V_{\text{ex}} \rangle$ , it is useful to define the dimensionless shape factor  $\langle v_{\text{ex}} \rangle$ , for a set of fractures with identical shapes, but

possibly different sizes

$$\langle v_{\text{ex}} \rangle = \frac{\langle V_{\text{ex}} \rangle}{\langle R \rangle \langle R^2 \rangle} \quad (28)$$

It can then be used to define two dimensionless densities, with different weightings of the fracture sizes

$$\begin{aligned} \rho'_{21} &= \rho \langle v_{\text{ex}} \rangle \langle R^2 \rangle \langle R \rangle = \rho \langle V_{\text{ex}} \rangle ; \\ \rho'_3 &= \rho \langle v_{\text{ex}} \rangle \langle R^3 \rangle . \end{aligned} \quad (29)$$

The subscripts are reminders of the statistical moments of  $R$  involved in each definition.  $\rho'_{21}$  is the generalization of  $\rho'$  for monodisperse networks, since it can be shown that it is still equal to the mean number of intersections per fracture [2]. Both  $\rho'_{21}$  and  $\rho'_3$  reduce of course to  $\rho'$  in case of equal-sized fractures.

The main tools required to study the percolation of polydisperse networks model are similar to the ones described in Sect. “Methods”.

For given values of the parameters, the probability  $\Pi$  of having a percolating cluster which spans the cell along the  $x$ -direction is derived from  $N_r$  random realizations of the system; then, the value  $\rho'_c$  for which  $\Pi = 0.5$  is estimated.  $\Pi$  and  $\rho'_c$  depend on several parameters as summarized by the formulae

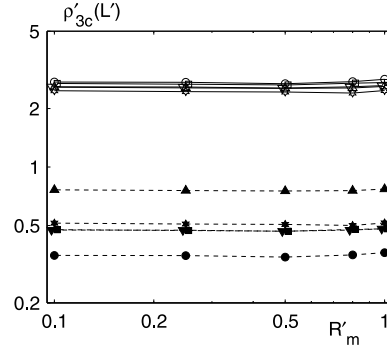
$$\Pi(R'_m, L', a, S, \rho') , \quad \rho'_c(R'_m, L', a, S) \quad (30)$$

where  $\rho'$  denotes any one of the dimensionless densities defined in (29). For brevity, they will be often written as  $\Pi(L', \rho')$  and  $\rho'_c(L')$ .

In practice,  $\Pi(L', \rho')$  was evaluated from sets of 500 realizations, for about 10 values of the network density, evenly distributed in a range where  $\Pi$  varies from 0.05 to 0.95. Since there is a correspondance between  $\rho'_{21}$  and  $\rho'_3$ , for given values of  $S$ ,  $a$  and  $R'_m$ , the same data sets can be used to determine  $\rho'_{21c}(L)$  and  $\rho'_{3c}(L)$ . The 95% confidence interval is estimated to be about  $\pm 0.04$  in terms of  $\rho'_{3c}(L)$ .

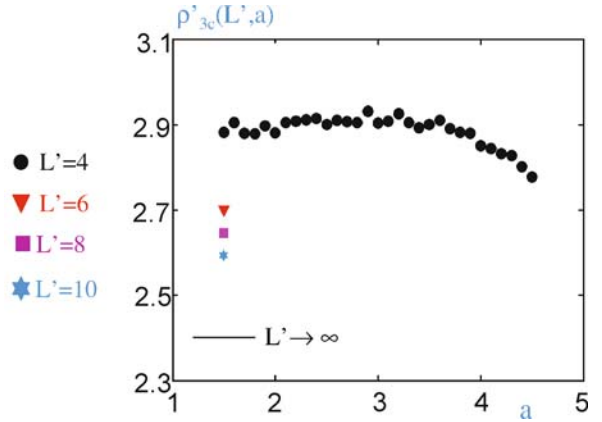
The influence on  $\rho'_c$  of the four parameters in Eq. (30) was systematically studied in [24]. We only state here the main result, which is that in the range  $1.5 \leq a \leq 4$ ,  $R'_m \ll L$  and for (almost) any fracture shape,  $\rho'_c$  depends only on the domain size, and that in the limit of infinite domains, a unique value of  $\rho'_c(\infty)$  applies in all cases. The independence on the various parameters is illustrated in the following examples.

In the example of Fig. 4,  $L'$  and  $a$  are kept constant, but the range of size and the fracture shapes varied. The networks contain hexagons, squares or triangles, or mixtures of hexagons with triangles or rectangles with a four



Percolation, and Faults and Fractures in Rock, Figure 4

The percolation thresholds  $\rho'_{3c}$  (open symbols, solid lines) and  $\rho_c \langle R^3 \rangle$  (black symbols, broken lines) for networks with  $L' = 6$  and  $a = 1.5$  for regular hexagons ( $\circ$ ), squares ( $\square$ ), triangles ( $\triangle$ ), mixture of hexagons and triangles, 50%–50% ( $\nabla$ ), and mixture of hexagons and rectangles with aspect ratio 4, 50%–50% ( $\star$ )



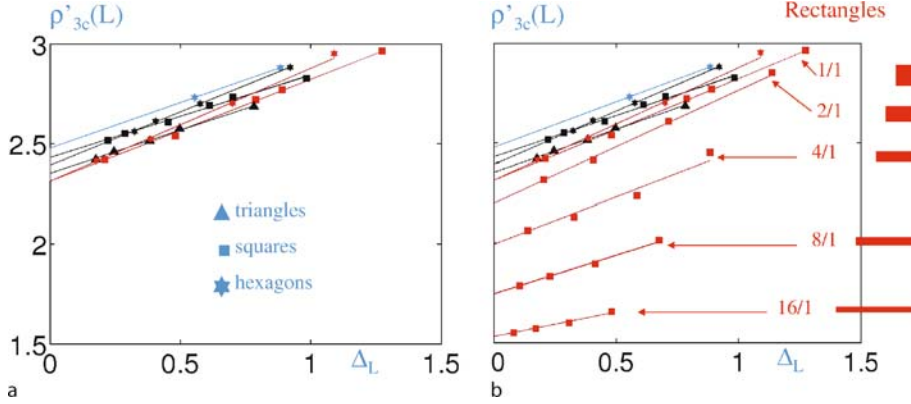
Percolation, and Faults and Fractures in Rock, Figure 5

The percolation threshold  $\rho'_{3c}(L', a)$  for networks of hexagonal fractures with  $R'_m = 0.1$ , versus the exponent  $a$ , for various domain sizes  $L$ . The lower line is the extrapolation of the data for  $a = 1.5$  when  $L'$  tends to infinity

to one aspect ratio. The upper set of curves shows that  $\rho'_c$  is indeed independent of  $R'_m$  and  $S$ . Note that the rightmost points are actually monodisperse networks. For comparison, the thresholds  $\rho_c \langle R^3 \rangle$ , which do not include the shape factor  $\langle v_{\text{ex}} \rangle$  (see Eq. 29), are also shown in the same figure and they are clearly much more scattered. It is the incorporation of  $\langle v_{\text{ex}} \rangle$  in the definition of  $\rho'_3$  which unifies the results for the different shapes.

Conversely, the fracture shape (hexagonal) and the range of size ( $R'_m = 0.1$ ) are kept constant in the example of Fig. 5, whereas the exponent  $a$  and the domain size  $L$  are varied. It is seen that  $\rho'_c$  does not vary when  $a$  ranges from 1.5 to 4. However, a definite dependence on the domain





**Percolation, and Faults and Fractures in Rock, Figure 6**

The percolation threshold  $\rho'_{3c}(L')$  for mono- or polydisperse networks of fractures with various shapes, versus the width  $\Delta_L$  of the percolation transition. In a, the fractures are hexagons, squares or triangles.  $\rho'_{3c}(\infty)$  is the extrapolation for  $\Delta_L \rightarrow 0$ , which falls in the range of Eq. 31. Data for monodisperse networks of rectangles with aspect ratios from 1 to 16 are added in b

size is observed, which corresponds to the well known finite size effects.

The data for increasing  $L'$  can be extrapolated for infinite systems by use of a classical technique. The combination of (23) and (24) shows that  $\rho'_c(L) - \rho'_c(\infty)$  is proportional to the width  $\Delta_L$  of the percolation transition zone. Hence,  $\rho'_c(\infty)$  can be read on the vertical axis of the plot of  $\rho'_c(L)$  versus  $\Delta_L$ , which is shown in Fig. 6. The data for many cases, including various fracture shapes in monodisperse and polydisperse networks, are gathered in Fig. 6a. In all cases, the extrapolated values  $\rho'_{3c}(\infty)$  fall in the narrow range

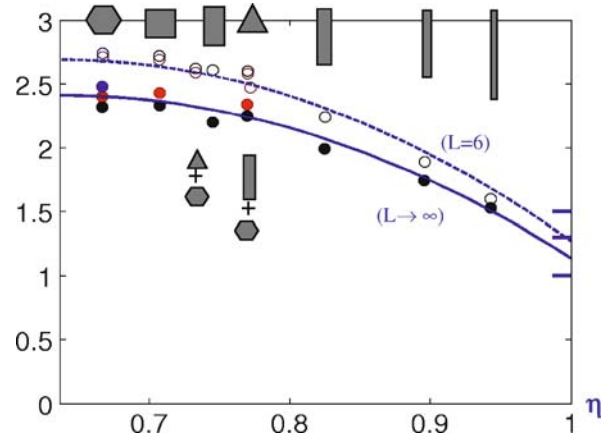
$$\rho'_{3c}(R'_m, a, S, L' \rightarrow \infty) = \rho'_{3c}(\infty) \approx 2.4 \pm 0.1. \quad (31)$$

This applies for a variety of shapes, as well as for mixtures of fractures with different shapes (see Fig. 4).

However, when the polygons become elongated,  $\rho'_{3c}(L')$  varies with the aspect ratio. Data for rectangles with aspect ratios  $A_r$  up to 16 are shown in Fig. 6b. It appears that  $\rho'_{3c}(L')$  decreases significantly when  $A_r$  increases.

This can be taken into account by using the shape factor  $\eta = 4R/P$  of the fractures. This ratio is minimum for disks, with  $\eta = 2/\pi \approx 0.637$ , and it increases up to one when the shape deviates from circularity. It turns out that a quadratic correction in terms of  $\eta$  is very successful for the representation of the data for very different and irregular fracture shapes.

All the thresholds obtained in cells with  $L' = 6$  and mono- or polydisperse size distributions with  $a = 1.5$  or 2 and  $R_m = 0.1$  are plotted in Fig. 7 as functions of  $\eta$ . This includes networks of hexagons, squares, triangles, mix-



**Percolation, and Faults and Fractures in Rock, Figure 7**

The percolation thresholds  $\rho'_{3c}(L' = 6)$  and  $\rho'_{3c}(\infty)$  for a variety of fracture shapes and size distributions, in comparison with the expressions (32), (33). The marks on the right are the predictions of [15,16,31] for infinitely elongated objects. The fracture shapes are indicated by the icons above or below the data points

tures of hexagons with rectangles or triangles, and rectangles with  $h/w$  up to 16. The data are well fitted by the expression

$$\rho'_{3c}(L') = 2.69 \left[ 1 - 4 \left( \eta - \frac{2}{\pi} \right)^2 \right] \quad (L' = 6). \quad (32)$$

The extrapolated data for infinite systems are also presented in Fig. 7, in comparison with the corrected version of Eq. (31),

$$\rho'_{3c}(\infty) = 2.41 \left[ 1 - 4 \left( \eta - \frac{2}{\pi} \right)^2 \right]. \quad (33)$$

In both cases, the deviations never exceed  $\pm 0.1$ . The corrective term becomes significant, i. e., larger than the error bar in (31), when  $\eta > 3/4$ , which corresponds for rectangles to aspect ratios larger than 2.

It can be noted that Eq. (33) predicts a threshold value 1.14 when  $h/w$  tends to infinity (i. e., when  $\eta \rightarrow 1$ ), which is in the range of the predictions 1.5 for prolate ellipsoids [16], 1 for capped cylinders [15] and 1.3 for elongated prisms [31], in the limit of infinite slenderness.

### Determination of the Dimensionless Density from Experimental Data

Since percolation properties are controlled by the dimensionless density  $\rho'$ , it is theoretically and practically important to derive estimations of  $\rho'$  from field data. In most cases, these data are based on 1D and 2D measurements of fracture traces along boreholes or on exposed outcrops which necessitate extrapolation by stereological techniques to 3D. Such extrapolations have already been made for specific fracture shapes by Warburton [39,40], Piggott [26], Berkowitz and Adler [8] and Sisavath et al. [36] (see also the references therein).

Our general methodology which is detailed in [38] can be illustrated by the intersection of a family of convex fractures with a line of length  $L$  which is parallel to the unit vector  $\mathbf{p}$ . Consider a fracture of surface  $A$ , of normal  $\mathbf{n}$  and of in-plane orientation  $\omega$ ; this object does not intersect the line when its center is located out of a surface of area  $A$ . Since this is valid for any in-plane orientation, the excluded volume of the line and of the surface is equal to  $AL|\mathbf{p} \cdot \mathbf{n}|$ . Hence, the average number of intersections  $\langle n_I \rangle$  per unit length between such a line and an isotropic network of a monodisperse family of fractures is

$$\langle n_I \rangle = \frac{1}{2} A \rho. \quad (34)$$

Of course, the major interests of this formula are that it does not depend on the precise shape  $S$  of the fractures and that  $\rho$  can be deduced from  $n_I$  and  $A$ . However, it depends in a crucial way on the convexity of the fractures.

### Isotropic Networks

In order to derive the average number of intersections  $\Sigma_t$  of a family of convex fractures  $\mathcal{F}(R)$  with a plane  $\Pi$  per unit area of the plane, define in  $\Pi$  a large convex region  $\mathcal{R}$  of area  $\mathcal{A}$  and perimeter  $\mathcal{P}$ . The excluded volume of  $\mathcal{F}(R)$  and  $\mathcal{R}$  is thus given by (8). The number of intersections  $d\Sigma_t$  of the fractures of size ranging from  $R$  to  $R + dR$  is proportional to the volumetric density of such fractures

multiplied by the excluded volume of  $\mathcal{F}(R)$  and  $\mathcal{R}$  as expressed by (8); when  $\mathcal{A} \rightarrow \infty$ ,  $\mathcal{A} \gg \mathcal{P}$ ; therefore,

$$d\Sigma_t(R) \rightarrow \frac{1}{4} \rho P(R) n(R) dR \quad \text{when } \mathcal{A} \rightarrow \infty. \quad (35)$$

This relation can be averaged over the sizes  $R$

$$\Sigma_t = \int d\Sigma_t(R) = \frac{1}{4} \rho \langle P \rangle. \quad (36)$$

The intersections of the fractures with a plane are called *traces* or *chords*. Let  $c$  be the length of a trace as illustrated in Fig. 1c. Such an intersection of length  $c(z, \mathbf{n}, \omega)$  exists if the vertical coordinate  $z$  of the center verifies

$$z_m(\mathbf{n}, \omega) \leq z \leq z_M(\mathbf{n}, \omega). \quad (37)$$

For a given fracture of size  $R$ , the average trace length  $\langle c \rangle_R$  when the intersection exists, can be expressed as

$$\langle c \rangle_R = \frac{\int d\omega \int d\mathbf{n} \int_{z_m}^{z_M} c dz}{\int d\omega \int d\mathbf{n} \int_{z_m}^{z_M} dz}. \quad (38)$$

Surprisingly, the numerator  $N_R$  of this fraction is easier to evaluate than its denominator  $D_R$ . The most internal integral  $\int_{z_m}^{z_M} c dz$  is equal to the area  $A$  of the fracture projected onto the plane perpendicular to  $\Pi$  which contains the trace, i. e.,  $A \sin \theta$ . Therefore,  $N_R$  is equal to  $\pi^3 A$ . The derivation of  $D_R$  is slightly more involved Santalo [34]; it is proportional to the integral of the Feret (or caliper) diameter over  $\omega$ . Frenet formulae are used to express this integral. Finally,

$$\langle c \rangle_R = \pi \frac{A(R)}{P(R)}. \quad (39)$$

For polydisperse fractures, the overall average  $\langle c \rangle$  is given by

$$\langle c \rangle = \frac{\int dR \Sigma_t(R) \langle c \rangle_R}{\int dR \Sigma_t(R)} = \pi \frac{\langle A \rangle}{\langle P \rangle} \quad (40)$$

a formula which is again an obvious generalization of the disk formula (cf. (24a) of Berkowitz and Adler [8]).

The density of trace intersections  $\Sigma_p$  is defined as the number per unit surface in the observation plane of the points which are intersections of traces. Since the fractures are randomly oriented and distributed in space, the same properties are valid for the traces. Moreover, as a trivial extension of the concept of excluded volume, the excluded surface  $S_{ex}$  of two traces of random orientations and of lengths  $c_1$  and  $c_2$  is equal to (cf. [2])

$$S_{ex} = \frac{2}{\pi} c_1 c_2. \quad (41)$$

Let  $\sigma_t(R, c)dc dR$  be the surface density of traces of length  $c$  ranging from  $c$  to  $c + dc$ , for the fractures of size  $R$  ranging from  $R$  to  $R + dR$ . Hence, the surface density of intersections of traces  $c_1$  corresponding to fractures of size  $R_1$  and of traces  $c_2$  corresponding to fractures of size  $R_2$  is

$$\sigma = \frac{1}{2} \sigma_t(R_1, c_1) \sigma_t(R_2, c_2) \frac{2}{\pi} c_1 c_2. \quad (42)$$

As a direct consequence

$$\Sigma_p = \iiint \sigma dc_1 dR_1 dc_2 dR_2. \quad (43)$$

This last expression can be split into a product of integrals since the populations 1 and 2 are independent. According to (35), (36), (43),  $\Sigma_p$  can be expressed as

$$\Sigma_p = \frac{1}{\pi} \frac{\pi^2}{16} \rho^2 \langle A \rangle^2 = \frac{\pi}{16} \rho^2 \langle A \rangle^2. \quad (44)$$

### Extensions

Let us now examine various possible extensions of the previous formulae.

The precise shape  $S$  of the fractures is never taken into account. Therefore, all the previous formulae are valid whatever the mixture of shapes  $S$ .

For anisotropic networks, the normal vector  $\mathbf{n}$  is not uniformly distributed over the unit sphere. Let  $\theta$  and  $\varphi$  be the two polar angles of  $\mathbf{n}$  (cf. Fig. 1c); the probability that the end of  $\mathbf{n}$  for fractures of sizes in the interval  $[R, R + dR]$  belongs to the interval  $[\theta, \theta + d\theta] \times [\varphi, \varphi + d\varphi]$  is given by  $\rho n(R, \mathbf{n}) d\theta d\varphi dR$ . The statistical average  $\langle \cdot \rangle$  can be calculated with this differential element.

The first quantity which can be easily generalized is  $\langle n_I \rangle$  (cf. (34))

$$\langle n_I \rangle = \rho \iiint n(R, \mathbf{n}) A \cos \theta d\theta d\varphi dR = \rho \langle A | \mathbf{p} \cdot \mathbf{n} | \rangle. \quad (45)$$

The other generalized formulae are summarized in Table 2.  $\alpha$  is the angle between the normal  $\mathbf{v}$  to the plane  $\Pi$  and  $\mathbf{n}$ ; in most cases, by choosing the  $z$ -axis perpendicular to  $\Pi$ ,  $\alpha$  is equal to  $\theta$ ;  $\beta_{12}$  is the angle between the normals  $\mathbf{n}_1$  and  $\mathbf{n}_2$  to the two fractures 1 and 2.

These formulae can be specialized to networks of subvertical fractures with a horizontal observation plane  $\Pi$ . Then,  $\alpha$  is equal to  $\pi/2$  and  $\beta_{12}$  is equal to the angles between the two traces in  $\Pi$ . Such networks can be either isotropic (i. e., the directions of the traces in  $\Pi$  are isotropic), or anisotropic. The corresponding results are detailed in Table 2.

### Percolation, and Faults and Fractures in Rock, Table 2

The major relations for the various kinds of networks.  $\mathcal{B}_{12} = \langle A_1 A_2 | \sin \beta_{12} | \rangle$

	Isotropic 3D	Anisotropic 3D	Subvertical isotropic	Subvertical anisotropic
$\langle n_I \rangle$	$\frac{1}{2} \rho \langle A \rangle$	$\rho \langle A   \mathbf{p} \cdot \mathbf{n}   \rangle$	$\frac{2}{\pi} \rho \langle A \rangle$	$\rho \langle A   \mathbf{p} \cdot \mathbf{n}   \rangle$
$\Sigma_t$	$\frac{1}{4} \rho \langle P \rangle$	$\frac{\rho}{\pi} \langle   \sin \alpha   P \rangle$	$\frac{\rho}{\pi} \langle P \rangle$	$\frac{\rho}{\pi} \langle P \rangle$
$\langle c \rangle$	$\pi \frac{\langle A \rangle}{\langle P \rangle}$	$\pi \frac{\langle A   \sin \alpha   \rangle}{\langle P   \sin \alpha   \rangle}$	$\pi \frac{\langle A \rangle}{\langle P \rangle}$	$\pi \frac{\langle A \rangle}{\langle P \rangle}$
$\Sigma_p$	$\frac{\pi}{16} \rho^2 \langle A \rangle^2$	$\frac{1}{2} \rho^2 \mathcal{A}_{12}$	$\frac{1}{\pi} \rho^2 \langle A \rangle^2$	$\frac{\rho^2}{2} \mathcal{B}_{12}$

### Discussion

**Discrete Families of Fractures** In many practical cases, the fractures are perpendicular to a finite set of normals  $\{\mathbf{n}_i; i = 1, \dots, m\}$  with probabilities  $\{n(R, \mathbf{n}_i); i = 1, \dots, m\}$ . The integrals over  $d\theta d\varphi$  are thus replaced by the following summation for a function  $f(R, \mathbf{n}_i)$

$$\rho \sum_{i=1}^m n(R, \mathbf{n}_i) f(R, \mathbf{n}_i). \quad (46)$$

**Practical Use of the Formulae** The major interest of the formulae summarized in Table 2 is to try to use them to derive the macroscopic quantities  $\rho$ ,  $\langle A \rangle$  and  $\langle P \rangle$ . It is easy (and frustrating) to realize that only two of these quantities can be obtained. For instance, (40) implies that  $\langle A \rangle = \pi^{-1} \langle P \rangle \langle c \rangle$ ; from (36),  $\langle P \rangle = 4\rho^{-1} \Sigma_t$ ; therefore  $\langle A \rangle = 4\pi^{-1} \rho^{-1} \Sigma_t \langle c \rangle$ . When these expressions are introduced into (34) or (44), one obtains that the three following ratios should be equal to one

$$\kappa_1 = \frac{\pi}{2} \frac{\langle n_I \rangle}{\Sigma_t \langle c \rangle}, \quad \kappa_2 = \frac{\pi \Sigma_p}{\Sigma_t^2 \langle c \rangle^2}, \quad \kappa_3 = \frac{\pi}{4} \frac{\langle n_I \rangle^2}{\Sigma_p}. \quad (47)$$

The third relation is derived by eliminating  $\Sigma_t \langle c \rangle$  between  $\kappa_1$  and  $\kappa_2$ . These relations provide consistency relations between the data, but not  $\rho$ .

In other words, only two of the three quantities  $\rho$ ,  $\langle A \rangle$  and  $\langle P \rangle$  can be simultaneously derived from the average measured data. Note also that  $\kappa_1$  is insensitive to the spatial organization, and that this is not true for  $\kappa_2$  and  $\kappa_3$  which depend on trace intersections.

One can go further if some geometrical information is available which could be  $\langle V_{ex} \rangle$ . Here, we shall use a shape factor  $\eta$  which is defined as  $\langle A \rangle \langle P \rangle^{-2}$ . For 3D isotropic networks, this expression can be combined to (40) and to (36) to yield  $\langle A \rangle$ ,  $\langle P \rangle$  and  $\rho$

$$\langle P \rangle = \frac{\langle c \rangle}{\pi \eta}, \quad \langle A \rangle = \frac{\langle c \rangle^2}{\pi^2 \eta}, \quad \rho = 4\pi \eta \frac{\Sigma_t}{\langle c \rangle} \quad (48)$$

or for a set of fractures normal to  $\mathbf{n}_i$

$$\rho_i = \frac{\pi^2 \eta_i}{|\sin \alpha_i| \langle c \rangle_i} \frac{\Sigma_{ti}}{\langle c \rangle_i} \quad (49)$$

There are many equivalent ways to derive  $\rho$ . The choice of the adequate formula depends on the available data. Note that formulae which contains  $\Sigma_p$  cannot be applied to families of parallel fractures.

When  $\rho$  and therefore  $\langle V_{ex} \rangle$  (cf. (8)) are known by one way or another, one can derive the dimensionless density  $\rho' = \rho \langle V_{ex} \rangle$  for isotropic and anisotropic networks

$$\rho' = \frac{\rho}{\pi} \langle (A_1 P_2 + A_2 P_1) |\sin \beta_{12}| \rangle \quad (50)$$

$$\rho' = \frac{\rho}{2} \langle A \rangle \langle P \rangle \quad (3d); \quad \rho' = \frac{4\rho}{\pi^2} \langle A \rangle \langle P \rangle \quad (2d). \quad (51)$$

Then, if the fracture network is not too polydisperse, one can use a classical mean field argument and approximate its properties by the properties of a monodisperse network of density  $\rho'$ .

## Applications

Several applications have already been made of the previous methodology and they can be summarized as follows. [36] showed that when data relative to fractures are collected along a line (e.g. a road or a well), estimations can be given to the major geometrical properties of the corresponding fracture networks, such as the volumetric density of fractures and their percolation character. [38] used the two dimensional maps obtained by [25] for subvertical fractures. Among other results, some of the consistency relationships (47) are well verified by these data. As previously,  $\rho'$  is estimated.

Finally [17], reconstructed a three-dimensional fracture network in a granite block from a series of experimental serial sections provided by [21]. It was visualized and its most important geometrical characteristics were studied. Though the network mostly consists of two families of fractures, it is interesting to note that a simple model of randomly oriented, monodisperse hexagons often yields a good order of magnitude for the various geometrical properties, which have been measured on the real block.

## Role of the Dimensionless Density in Other Geometrical Properties and Permeability

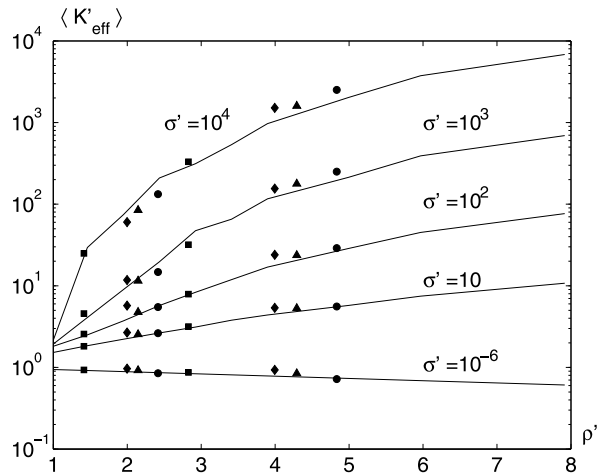
Though this article is focused on percolation properties, it is important to notice that the dimensionless densities

which were introduced, play a crucial role in other properties as well. [18] studied two main other geometrical properties for monodisperse networks. Fracture networks partition the solid space into blocks; the block density is denoted by  $\rho_b$ . One can introduce the cyclomatic number of the graph  $\Gamma_1$  which is the number of independent cycles of this graph, and more precisely the number of cycles  $\bar{\beta}_1$  per unit volume. [18] showed that  $\rho_b$  and  $\bar{\beta}_1$  when made dimensionless by the excluded volume are independent of the fracture shapes.

Similar properties are found for the macroscopic permeability of fracture networks [20] and of fractured porous media whether they are monodisperse [10] or polydisperse [23]. In a series of contributions, the corresponding dimensionless quantities were shown to depend only on the dimensionless density  $\rho'$ . This is illustrated in Fig. 8. The porous medium has a local permeability  $K_m$  and the monodisperse fractures a conductivity  $\sigma$ . The macroscopic permeability of this medium is denoted by  $K_{eff}$ . Dimensionless quantities denoted by primes can be defined as

$$\sigma = R K_m \sigma', \quad K_{eff} = K_m K'_{eff}. \quad (52)$$

Figure 8 shows that the average macroscopic permeability  $\langle K'_{eff} \rangle$  does not depend significantly on the fracture shape. The two major parameters are  $\rho'$  and  $\sigma'$ .



Percolation, and Faults and Fractures in Rock, Figure 8

Statistical averages of the permeability  $\langle K'_{eff} \rangle$  for samples containing  $N_f=16$  or 32 fractures, with 4-, 6- or 20-gonal shapes, as functions of the network density  $\rho'$  and of the fracture conductivity  $\sigma'$ . The cell size is  $L = 4R$ . Data are for squares ( $\square$ ), rectangles with aspect ratios two to one ( $\Delta$ ) or four to one ( $\diamond$ ), hexagons (lines) and icosagons ( $\circ$ )



## Future Directions

The percolation properties of networks of random and convex plane fractures are successfully addressed by means of the excluded volume. Many important dimensionless properties of isotropic fracture networks only depend on the dimensionless density of fractures and not on the fracture shapes and sizes which represents a significant simplification.

These properties are not specific of fractures present in rocks and the same methodology can be applied for any other fracture system whatever the characteristic sizes and the nature of the material where it occurs.

These results should be generalized in several directions. In most cases, real fractures are not isotropically oriented and this feature should be incorporated in the next studies on this subject. The same is true for the homogeneous character of the network.

## Bibliography

### Primary Literature

- Adler PM (1992) Porous Media: Geometry and Transports. Butterworth/Heinemann, Stoneham
- Adler PM, Thovert J-F (1999) Fractures and fracture networks. Kluwer Academic Publishers, Dordrecht
- Alon U, Balberg I, Drory A (1991) New, heuristic, percolation criterion for continuum systems. *Phys Rev Lett* 66:2879–2882
- Balberg I (1985) Universal percolation threshold limits in the continuum. *Phys Rev* B31:4053–4055
- Balberg I (1987) Recent developments in continuum percolation. *Phil Mag* B56:991–1003
- Balberg I, Anderson CH, Alexander S, Wagner N (1984) Excluded volume and its relation to the onset of percolation. *Phys Rev* B30:3933–3943
- Barenblatt GI, Zheltov IP, Kochina IN (1960) Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks. *Soviet Appl Math Mech (PMM)* 24:852–864
- Berkowitz B, Adler PM (1998) Stereological analysis of fracture network structure in geological formations. *J Geophys Res* B103:15339–15360
- Berkowitz B, Ewing RP (1998) Percolation theory and network modeling applications in soil physics. *Survey Geophys* 19:23–72
- Bogdanov I, Mourzenko VV, Thovert J-F, Adler PM (2003) Effective permeability of fractured porous media in steady state flow. *Water Resour Res* 39. doi:10.1029/2001WR000756
- Bour O, Davy P (1997) Connectivity of random fault networks following a power law fault length distribution. *Water Resour Res* 33:1567–1583
- Charlaix E, Guyon E, Rivier N (1984) A criterion for percolation threshold in a random array of plates. *Solid State Commun* 50:999–1002
- Conrad F, Jacquin C (1973) Représentation d'un réseau bidimensionnel de fractures par un modèle probabiliste. Application au calcul des grandeurs géométriques des blocs matriciels. *Rev IFP* 28:843–890
- Drory A, Berkowitz B, Parisi G, Balberg I (1997) Theory of continuum percolation. III. Low-density expansion. *Phys Rev* E56:1379–1395
- Florian R, Neda Z (2001) Improved percolation thresholds for rods in three-dimensional boxes. oai:arXiv.org:cond-mat/0110067
- Garboczi EJ, Snyder KA, Douglas JF, Thorpe MF (1995) Geometrical percolation threshold of overlapping ellipsoids. *Phys Rev* E 52:819–828
- Gonzalez Garcia R, Huseby O, Thovert J-F, Ledésert B, Adler PM (2000) Three-dimensional characterization of fractured granite and transport properties. *J Geophys Res* 105(B)21387–21401
- Huseby O, Thovert J-F, Adler PM (1997) Geometry and topology of fracture systems. *J Phys A* 30:1415–1444
- Ishihara A (1950) Determination of molecular shape by osmotic measurement. *J Chem Phys* 18:1446–1449
- Koudina N, Gonzalez Garcia R, Thovert J-F, Adler PM (1998) Permeability of three-dimensional fracture networks. *Phys Rev* E 57:4466–4479
- Ledésert B, Dubois J, Velde B, Meunier A, Genter A, Badri A (1993) Geometrical and fractal analysis of a three-dimensional hydrothermal vein network in a fractured granite. *J Volcanol Geotherm Res* 56:267–280
- Long JCS, Remer JS, Wilson CR, Witherspoon PA (1982) Porous media equivalents for networks of discontinuous fractures. *Water Resour Res* 18:645–658
- Mourzenko V, Thovert J-F, Adler PM (2004) Macroscopic permeability of three dimensional fracture networks with power law size distribution. *Phys Rev* E 69:066307
- Mourzenko V, Thovert J-F, Adler PM (2004) Percolation of three-dimensional fracture networks with power-law size distribution. *Phys Rev* E 72:036103
- Odling NE (1997) Scaling and connectivity of joint systems in sandstones from western Norway. *J Struct Geol* 19:1257–1271
- Piggott AR (1997) Fractal relations for the diameter and trace length of disc-shaped fractures. *J Geophys Res* 102(B):18121–18125
- Pike GE, Seager CH (1974) Percolation and conductivity: A computer study. I. *Phys Rev* B10:1421–1434
- Rivier N, Guyon E, Charlaix E (1985) A geometrical approach to percolation through random fractured rocks. *Geol Mag* 122:157–162
- Robinson PC (1983) Connectivity of fracture systems - A percolation theory approach. *J Phys* A16:605–614
- Robinson PC (1984) Numerical calculations of critical densities for lines and planes. *J Phys* A17:2823–2830
- Saar MO, Manga M (2002) Continuum percolation for randomly oriented soft-core prisms. *Phys Rev* E 65:056131
- Sahimi M (1995) Flow and transport in porous media and fractured rocks. VCH, Weinheim
- Sahimi M, Yortsos TL (1990) Applications of Fractal Geometry to Porous Media: A review. Society of Petroleum Engineers. Paper 20476
- Santalo LA (1943) Sobre la distribución probable de corpusculos en un cuerpo, deducida de la distribución en sus secciones y problema analogos. *Rev Unión Mat Argent* 9:145–164
- Sher H, Zallen R (1970) Critical density in percolation processes. *J Chem Phys* 53:3759–3761
- Sisavath S, Mourzenko V, Genthon P, Thovert J-F, Adler PM (2004) Geometry, percolation and transport properties of frac-

- ture networks derived from line data. *Geophys J Int line-break* 157:917–934
37. Stauffer D, Aharony A (1994) *Introduction to Percolation Theory*, 2nd edn. Taylor and Francis, Bristol
  38. Thovet J-F, Adler PM (2005) Trace analysis for fracture networks of any convex shape. *Geophys Res Lett* 31:L22502
  39. Warburton PM (1980a) A stereological interpretation of joint trace data. *Int J Rock Mech Min Sci Geomech Abstr* 17:181–190
  40. Warburton PM (1980b) Stereological interpretation of joint trace data: Influence of joint shape and implication for geological surveys. *Int J Rock Mech Min Sci Geomech Abstr* 17:305–316

### Books and Reviews

- Bear J, Tsang C-F, de Marsily G (1993) *Flow and contaminant transport in fractured rock*. Academic Press, San Diego
- Myer LR, Tsang CF, Cook NGW, Goodman RE (1995) *Fractured and jointed rock masses*. Balkema, Rotterdam
- van Golf-Racht TD (1982) *Fundamentals of fractured reservoir engineering*. Developments in Petroleum Science, vol 12. Elsevier, Amsterdam

## Percolation, Introduction to

MUHAMMAD SAHIMI

Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, USA

Why is percolation theory relevant to the analysis of complex systems? The question can be answered only if we first define what we mean by a complex system. I spent a large amount of time in vain, searching for a “clean,” generally-accepted definition of a complex system, until I finally realized that there are probably as many definitions as the number of scientists that deal with complex systems – there is not a clean *universal* definition of a complex system.

But, at the very least, we can agree that a complex system consists of a large number of interacting components, or parts. The interactions may be short- or long-ranged, and may or may not change with time. One type of such interactions is the *connectivity*, the way the components or parts of a complex system are connected with each other. Clearly, if the components or parts of a complex system are not connected, they do not interact with each other, at least not directly. Now, if the connectivity of the components or parts of a complex system plays an important role on the macroscopic, or effective, properties of the system, then, percolation theory plays a prominent – and in fact a decisive – role in quantifying the effect of the connec-

tivity on the effective properties of that complex system, hence the inclusion of this section in the Encyclopedia.

Percolation processes are, in fact, the *opposite* of diffusion processes. In the latter case, the diffusant decides where to diffuse or move. The medium in which the diffusant is moving does not have any influence on the motion. This explains why many diffusion processes can be reduced to problems in essentially one-dimensional systems (all that matters is the distance  $r$  of the diffusant from the origin of its motion) which are, therefore, amenable to rigorous theoretical analysis and analytical solutions.

In a percolation process, on the other hand, it is the medium that decides where a species can go. Therefore, if the connectivity of the components or different parts of the medium – the complex system – is poor, the species cannot go far. If, on the other hand, the parts are well-connected, then the species is free to go almost anywhere. In political jargon, a complex system in which the connectivity of its different parts or components plays a decisive role in determining its effective properties – i. e., a percolation system – is like a corrupt society in which what matters is the connectivity to the powerful people! If one is well connected, one can advance rapidly; if not, there is little prospect for advancement. Since most media of interest are three- or at least two-dimensional systems, percolation problems are far more difficult to solve than the diffusion problems.

Why should the connectivity of a complex system be poor? Because natural complex systems are *disordered*. In fact, Nature, the most complex system that we know of, is disordered. Pure and geometrically perfect (periodic) systems are nowhere to be found, except perhaps in books and in our imaginations. One way that the disorder manifests itself is in the connectivity of the parts or components of a complex system. In some sectors of the system the parts are well connected, while in other sectors they are not. But, what matters most is the *overall* connectivity of the system, so that a given phenomenon can take place *across* the system.

An illuminating example is provided by porous materials. A piece of rock is a disordered porous medium: The shapes, sizes, and orientations of the pores are not identical, but vary greatly. If we attempt to characterize their statistical distributions, we find them to be broad. Thus, such a porous medium is a highly disordered and complex system. Clearly, the way the pores are connected plays a crucial role in flow of a fluid through the medium, which is why percolation is relevant to the description of fluid flow in disordered porous medium.

But, many of the man-made systems are also disordered, and the connectivity of their parts is a con-

trolling factor in determining their effective properties. For example, small molecules, or monomers, react, form chemical (covalent) bonds between themselves, and create a macromolecule. The structure of such a macromolecule is highly disordered. Clearly, the effective properties of the macromolecule is controlled by the connectivity of the monomers or the small molecules that formed it. If each monomer reacts with only two other monomers, it has a connectivity of two. The macromolecule has, therefore, no branches or loops. If, on the other hand, each monomer reacts with several others and has a higher connectivity, one obtains branched polymers and gels, the properties of which are completely different from those of linear polymers.

As another example, consider electrical conduction through a composite material which is a mixture of conducting and insulating phases. Assume that the two phases are randomly distributed in the composite. For simplicity, we model the composite material by a simple-cubic network in which each bond is either conducting with a finite conductivity, or insulating with zero conductivity. Suppose that we impose a voltage difference between two opposite faces of the network. The question then is: what fraction of the bonds must have a finite conductivity in order for the electrical current to flow through the material, so that it would have a nonzero macroscopic conductivity? This is clearly an important practical question, because its answer tells us, for example, what (volume) fraction of a composite material, such as carbon black composites that are used in many applications, must be conducting in order for the composite as a whole to be conducting.

If too many bonds are insulating, no macroscopic current will flow through the material, whereas for sufficiently large number of conducting bonds electrical current does flow in the material, so that its macroscopic effective conductivity is nonzero. If the fraction of the conducting bonds is  $p$ , then, there must be a minimum or critical value  $p_c$  of  $p$ , such that for  $p \leq p_c$  no electrical current would flow through the material and, therefore, the material as a whole is insulating, whereas for  $p > p_c$  the material becomes conducting.

The quantity  $p_c$  signals a phase transition: for  $p \leq p_c$  there is no sample-spanning path of conducting bonds, so that the material is macroscopically insulating – the conducting bonds are not macroscopically connected. For  $p > p_c$ , on the other hand, the system becomes macroscopically conducting – the conducting bonds are macroscopically connected. Hence,  $p_c$  is the point at which a *geometrical* phase transition from a disconnected to a connected system takes place. Percolation theory, then, quantifies the phase transition and its effect on the macro-

scopic properties of a complex system.  $p_c$  is called the *percolation threshold* of the medium.

Determination of the exact percolation thresholds of many 2D and all the 3D lattices remains an unsolved problem. John Wierman's article, [► Percolation Thresholds, Exact](#), describes and discusses the existing exact results for the percolation thresholds. Robert Ziff's article, [► Percolation Lattices, Efficient Simulation of Large](#), describes highly efficient numerical methods for estimating the percolation threshold and many other properties of percolation lattices, while Dietrich Stauffer's article, [► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation](#), describes the theoretical foundations for many important percolation properties near the percolation threshold.

The aforementioned articles describe percolation systems in which there is no correlations, and the complex system is completely random. However, disorder in many important heterogeneous materials is not completely random. There usually are correlations with an extent that may be finite but large. For example, in packing of solid particles, there are short-range correlations. Moreover, if the correlation function  $C(r)$  decays as  $r^{-d}$  or faster, where  $d$  is the Euclidean dimensionality of the system, then many properties of the system are very similar with those of random percolation. This is not totally unexpected because even in random percolation, as  $p$  decreases toward  $p_c$ , correlations begin to build up and, therefore, the introduction of any type of correlation with a range shorter than the percolation correlation length cannot change the properties fundamentally. In many other cases, e.g., in some disordered elastic materials, there are very strong correlations. The article by Antonio Coniglio and Annalisa Fierro, [► Correlated Percolation](#), describes the major differences between percolation in random and correlated systems.

A particular type of percolation model with extended correlations is known as the *bootstrap percolation*. In this problem the sites of a lattice are initially randomly occupied. Then, those sites that do not have at least  $Z_c$  nearest-neighbor occupied sites are removed (note that  $Z_c = 0$  is the usual random percolation). The interactions between the sites are short-ranged, but the correlations between them may build up as the distance between two occupied sites also increases. It now appears that bootstrap percolation possesses many unusual properties and, in fact, simulating and obtaining accurate estimates of this particular type of percolation systems have proven to be very difficult. The article by Paolo De Gregorio, Aonghus Lawlor, and Kenneth Dawson, [► Bootstrap Percolation](#), describes the progress in this important area of percolation prob-

lems, which appears to have applications to many important phenomena.

The example of conduction in composite materials that we described earlier also provided a hint on how complex percolation systems are modeled: They are represented by lattices or networks, such as the simple-cubic lattice. However, percolation in continua is also of great interest, since in most, if not all, of the practical applications one must deal with continuous systems. For example, continuum percolation is directly applicable to characterization and modeling of the morphology and effective transport properties of microemulsions, polymer blends, sintered materials, sol-gel transitions, and many other important practical problems. The article by Isaac Balberg, [► Continuum Percolation](#), describes the advances that have been made in understanding of the percolation effects in continuous systems.

Although percolation in lattice and continuous models has been studied extensively, in recent years another type of lattice model has attracted wide attention, as it appears to be applicable to a wide variety of phenomena, ranging from social dynamics to biological systems. These are lattices in which a small fraction of the sites – called the *hubs* – are highly connected, while a vast fraction of the remaining sites are connected only sparsely. Thus, the connectivity  $k$  of the nodes follows a statistical distribution. The connections are no longer necessarily between nearest-neighbor nodes. In fact, in such models the usual notion of nearest-neighbor nodes is not even useful or applicable. Some of such networks are *scale-free*. These are lattices in which the distribution of the nodal connectivities are of power-law type,  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  can be quite large. Percolation in such scale-free networks has very unusual properties. The article by Reuven Cohen and Shlomo Havlin, [► Percolation in Complex Networks](#), provides a review of this rapidly developing field. They describe how the concepts of percolation can be used to study not only the robustness and vulnerability of random networks, but also such problems as immunization and epidemic spreading in populations and computer networks, communication paths, and fragmentation in social networks.

Other major applications of percolation theory include modeling of transport in disordered materials, and in particular composite solids, and porous media. The transport processes that have been studied include fluid flow, diffusion, conduction, and deformation (stress transport), including computation of the elastic moduli. One important application of the percolation concepts is to two-phase fluid flow in porous media. In such problems, one fluid is injected into a porous medium – that is, it invades the

medium – in order to displace a second fluid already in the medium. A well-known example, practiced by the oil industry, is displacement of oil in an oil reservoir by water which is injected into the reservoir through some injection wells. Thus, this particular model is usually known as the invasion percolation. The article by Mark Knackstedt and Lincoln Paterson, [► Invasion Percolation](#), describes in detail how percolation is used to describe two-phase flow in porous media, although invasion percolation has proven to be relevant to many other phenomena. Other aspects of the application of percolation to problems that deal with fluid flow through porous media are described in the article by Peter King and Mohsen Masihi, [► Percolation in Porous Media](#).

The article by Barry Hughes, [► Conduction and Diffusion in Percolating Systems](#), provides a detailed description of diffusion and conduction in disordered and composite materials, including porous media, and presents a comprehensive account of the state-of-the-art of this important application of percolation to a problem of great practical importance.

Natural porous media are often fractured. For example, most oil reservoirs in the Middle East are fractured. The fractures often a large connected network, without which many oil reservoirs (such as those in Iran) could not produce any oil. Fractures also play an important role in fast transport and spreading of contaminants in groundwater aquifers. At the same time, natural porous media often contain large faults (one of the most famous of which, the San Andreas fault, is in California), which play the primary role in earthquakes. It now appears that percolation provides a powerful tool for modeling of the effect of the connectivity of fractures and faults on fluid flow and transport properties of rock, a highly complex set of phenomena. Fracture and faults also affect other important phenomena in rock, such as propagation of elastic and seismic waves. In their article, [► Percolation, and Faults and Fractures in Rock](#), Pierre Adler, Jean-Francois Thovert, and Valeri Mourzenko describe the recent progress in the application of percolation to this highly important set of problems.

Deformation and stress transport in disordered materials are important to a wide variety of phenomena in science and technology, ranging from elastic properties of solid materials, to viscoelastic properties of polymers and gels, and rigidity of biological materials (cells, proteins, bones, etc.). The application of percolation to modeling of such phenomena has proven to be highly successful and fruitful. The article by Phillip Duxbury, [► Elastic Percolation Networks](#), provides a comprehensive discussion of the subject, and describes the theoretical foundations and

computer simulation methods for stress transport in disordered percolation systems.

Two other articles expand on what Duxbury describes in his article. In his article, ► [Networks, Flexibility and Mobility in](#), Michael Thorpe describes recent advances on generalization of the percolation model, and its application to modeling of proteins and other biological materials. The question of the rigidity of such materials is addressed. An important and well-established application of percolation is to modeling of the rheology of polymers and gels, particularly in the vicinity of the gelation point. Several variants of the percolation models have been developed in order to address this important problem. The article by Muhammad Sahimi, ► [Percolation and Polymer Morphology and Rheology](#), describes the advances that have been made in this area.

Over the past three decades percolation theory has been applied to modeling of a wide variety of phenomena in disordered media and complex systems. It is impossible to describe and discuss all such applications. The applications that are described in this section of the Encyclopedia represent just the tip of the iceberg, but they do provide the reader with a good understanding of the power of percolation theory for describing a wide variety of important phenomena in complex systems. Have a good read!

## Percolation Lattices, Efficient Simulation of Large

ROBERT M. ZIFF

Department of Chemical Engineering,  
University of Michigan, Ann Arbor, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Cluster Identification and Growth](#)

[Hull Walks and Hull-Generating Walks](#)

[Gradient Percolation](#)

[The Microcanonical-Canonical Method](#)

[Other Numerical Techniques](#)

[Conductivity and Backbones](#)

[Conclusions](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Hull** The boundary of a percolation cluster, either internal or external.

**Accessible hull** The hull with pinched off “fjords” removed.

**Hull-generating walk** A way to generate the hull of a percolation cluster by a type of kinetic self-avoiding walk.

**Queue** A computer list construct in which the events are stored in such that the first in is the first out (also called “FIFO” or “breadth-first searching”).

**Stack** A computer list construct in which the events are stored in such a way that the last in is the first out (also called “LIFO” or “depth-first searching”).

**Recursion** A programming method in which a procedure calls itself, creating new local variables each time.

**Tree** A data structure in which points are connected in a tree-like structure with branches but no loops.

**Stochastic Loewner evolution (SLE)** A theoretical way to study conformally invariant random curves, including the hulls of percolation clusters, through a transformation of simple Brownian motion. Also called Schramm–Loewner Evolution.

**Leath algorithm** A technique where individual percolation clusters are “grown” from a seed by an epidemic type of process.

**Hoshen–Kopelman algorithm** A technique where a lattice (in 2d) is scanned one row at a time, and clusters are identified using information from the previous row only.

**Newman–Ziff algorithm** A way to efficiently generate microcanonical (fixed occupancy) states and from them to study all canonical (fixed  $p$ ) states.

### Definition of the Subject

Percolation is a simple model of the formation of long-range connectivity in random systems. While it can be solved exactly in a few cases of branched lattices, and while many results in two dimensions (2d) can be found exactly, most of the work in this field is intimately connected with computer simulation. Various algorithms have been developed over the years, and this article surveys some of them, especially related to cluster sizes and connectivity, and the hull.

### Introduction

Percolation was introduced by Flory in 1941 [13] for branched networks (polymers) and Broadbent and Hammersley in 1957 for lattice networks, and its study by computer simulation began just a few years later [77]. The



overall development of the percolation field in the ensuing years has been intimately connected with advances in simulation and computer algorithms. Specific computer algorithms allow optimal simulation of different aspects of the percolating system, and the results of these simulations have provided information and ideas for theoretical developments and further understanding of this fundamental problem. Recent advances have allowed, for some examples, the determination of numerical thresholds to very high precision, the demonstration (later confirmed by theory) that universality applies to crossing and excess cluster properties, and the universality of scaling functions and their resulting amplitude ratios.

The basic percolation system is a lattice with sites (vertices) and/or bonds (edges) occupied with a given probability  $p$ , and the computational problem is to identify and characterize the clusters of adjacently occupied sites (site percolation) or of sites connected by the bonds (bond percolation), and determine properties such as the size distribution, conductivity, and crossing. In principle, these are generally rather straightforward problems, but in practice, the challenge is to do things efficiently so that large systems can be simulated over many times to get good numerical significance.

In this article, we will describe and summarize several algorithms that have been developed to simulate percolation. Explicit fragments of programming in C are given. The emphasis of the article is on the computational methods, and we use some recent examples of application to illustrate them. We do not address the large number of advances that have been made recently in the percolation field mainly by mathematicians, based upon Stochastic Loewner Evolution (SLE) [61] and related methods. For a recent review, see [24].

Another area of recent interest related to percolation is in the study of networks, including Erdős-Rényi random graphs [11], scale-free and small world networks. For these models, percolation corresponds to the formation of a “giant component”. Because of the lack of loops over large distances, the percolation properties can generally be solved for analytically (i. e., [42,58]). Some of these results are closely related to percolation on the branch-free Bethe lattice, whose solution goes back to Flory and was analyzed in detail by Fisher and Essam [12]. However, in this review we will only consider regular lattices, and not the algorithms and results related to these systems.

### Cluster Identification and Growth

Consider a lattice, say square for simplicity, and suppose that the sites have been “populated” by being made “oc-

cupied” (OCC) with probability  $p$  and empty or “vacant” (VAC) with probability  $1 - p$ . (Here we are considering site percolation.) A basic problem is to identify the clusters, and to determine some property such as the size distribution  $n_s$  (the number of clusters of size  $s$ , divided by the total number of sites on the lattice) or whether crossing exists between two intervals on the boundary. Here we describe two neighbor-search methods to identify clusters: the depth-first or last-in, first-out (LIFO) method using recursion, and the breadth-first or first-in, first-out (FIFO) method using a queue.

In the following programs, we use a simple two-dimensional array `lat[x][y]` to represent a rectangular  $W \times H$  system with a square lattice. For many problems it is more efficient to use a one-dimensional array and add 1, -1,  $W$ , and  $-W$  for the four directions (in 2d, for example), using wrap-around at the end to form “helical” boundary conditions, which for a large system is practically equivalent to a periodic one. (The helicity adds a “twist” to the boundaries, which can have an effect on some of the properties when it is large enough [93]). Later on, in Sect. “The Microcanonical-Canonical Method”, we will give an example of a program which uses a one-dimensional array and also a neighbor array that can be used to program periodic boundary conditions precisely, with a bit more programming overhead however.

### Recursive Search (LIFO)

In the recursive search method, the lattice (say of dimensions  $W \times H$ ) is first scanned for new, unchecked sites:

```
for (xo = 0; xo < W; ++xo)
  for (yo = 0; yo < H; ++yo)
    if (lat[xo][yo] == OCC)          (P1)
      { lat[xo][yo] = TAGGED;
        FindNeighbors(xo,yo); }
```

where TAGGED means that the site has been checked and won't be checked again. Then the cluster belonging to that site is found using the `FindNeighbors` subroutine, which is given by

```
FindNeighbors(int x,y)
{ int dir;
  for (dir = 0; dir < 4; ++dir)
    { xp = x + dx[dir];
      yp = y + dy[dir];          (P2)
      if (lat[xp][yp] == OCC)
        { lat[xp][yp] = TAGGED;
          FindNeighbors(xp,yp); } }
```

When an OCC neighbor is found for the first time, it is checked by calling the same routine over again.

Here we used the four nearest-neighbor direction vectors  $dx[0] = 1$ ,  $dy[0] = 0$ ,  $dx[1] = 0$ ,  $dy[1] = 1$ ,  $dx[2] = -1$ ,  $dy[2] = 0$ ,  $dx[3] = 0$ ,  $dy[3] = -1$ . We have not dealt with the boundaries in this example. Open boundaries can be simulated by adding a perimeter of VAC sites; periodic boundary conditions can be simulated simply by writing  $lat[xp \& (W-1)][yp \& (H-1)]$  for  $lat[xp][yp]$ , where  $\&$  is the bit-wise “and” operation, if  $W$  and  $H$  are exactly powers of two. As the clusters are identified, the number of occupied sites can be counted, moments determined, etc. Crossing can be determined if a single cluster touches two given boundaries. For periodic b.c., one typically considers not crossing but wraparound.

The above `FindNeighbors` program uses recursion and the compiler stores unchecked  $xp, yp$  and the local  $dir$  in the stack; recursion uses a “last in, first out” (LIFO) or “depth-first” [71] method that can cause problems (exhaust the available stack memory) for very large clusters.

### Making a Queue (FIFO)

A more memory-efficient method to search for neighbors is to use a “queue” where one keeps a list of unchecked sites, and visits them in a first-in, first-out (FIFO) fashion. However, one must construct the list explicitly; the recursive method won’t do it. We need functions to put and remove coordinates from the queue; here we use `%define` statement functions (which creates in-line text) for efficiency:

```
%define PutOnQueue(X,Y) \
{ xlist[putindex] = X; \
  ylist[putindex] = Y; \
  ++putindex; }      (P3)
```

and

```
%define GetFromQueue(X,Y) \
{ X = xlist[getindex]; \
  Y = ylist[getindex]; \
  ++getindex; }      (P4)
```

where we start the simulation with  $getindex = putindex = 0$ . The way the above lists are written, they must be dimensioned to be as large as the largest possible cluster; however, by making the list size  $S$  exactly a power of two, the list can be shortened and “recycled” simply by writing  $xlist[putindex \& (S-1)]$  etc. The size  $S$  has only to be as large as the number of growth sites of a cluster, which grows as roughly the square root of the maximum size of the cluster. For example, for a lattice of size  $1024 \times 1024$ , it is more than suffi-

cient to make  $S = 4096$ . To test for an error in the recycled queue, the line `if (putindex == getindex) ...` can be added after `++putindex;` in (P3).

To make use of the queue, (P1) is kept the same except that the last line is replaced by `PutOnQueue(xo, yo)`, followed by the loop

```
do
{ GetFromQueue(x,y)
  for (dir = 0; dir < 4; ++dir)
  { xp = x + dx[dir];
    yp = y + dy[dir];
    if ((lat[xp][yp]) == OCC)
    { lat[xp][yp] = TAGGED;
      PutOnQueue(xp,yp); } }
  } while (getindex != putindex);      (P5)
```

When the two indices `putindex` and `getindex` are equal to each other, there are no more occupied sites to check and the search is complete.

Note that the lists `xlist` and `ylist` can also be used for the LIFO method by treating them as a stack rather than a queue – that is, by using only one index and to decrement the index when a set of coordinates is taken off the stack. This is a way to program the LIFO method without using the recursive feature of the C language.

### Generating Occupied Sites or Bonds as you go – the Leath Method

In the above algorithms, the sites or bonds were made occupied or vacant ahead of time. An alternative scheme is to start with all sites in an UNVISITED state, and make them OCC or VAC when they are first encountered. For example, in (P5), starting from the sixth line, we would write instead

```
if ((lat[xp][yp]) == UNVISITED)
  if (random() < prob)
  { lat[xp][yp] = OCC;
    PutOnQueue(xp,yp); }
else lat[xp][yp] = VAC;      (P6)
```

where `random()` is the random number generator and `prob` is the probability. Here, the label `TAGGED` is not needed, because a site is added to the queue as soon as it is made occupied. Before putting the first site  $xo, yo$  on the queue, one also has to call the random number generator to determine whether that site is occupied.

When this program is applied to the growth of a single cluster (that is, starting with just one  $xo$  and  $yo$ ), it is commonly called the Leath method, although it is not carried out in the same way as in [37], where larger diamond-shaped regions of the lattice around the cluster are

successively probed. In fact, the idea of looking at clusters wetting a single site was carried out in the earliest simulations of percolation [77].

The above growth scheme works particularly nicely in the case of bond percolation, especially when the clusters are characterized solely by the sites they connect, and not by the number or arrangement of the bonds that connect them. In this case, (P6) is kept exactly the same except the last line “else lat[xp][yp] = VAC;” is removed. This is because one can consider bond percolation as a spreading or epidemic process [18] from site to site along the occupied bonds; if a bond is not occupied, the fluid will not spread to the next site, but that does not preclude fluid from another cluster to visit that neighboring unvisited site. Sites with no bonds are considered to be clusters of size  $s = 1$ . Whether a bond is made occupied or not, that bond will never be considered again, so its state does not have to be remembered. Not all bonds of a cluster are generated in this method – for example, the fourth bond in a simple square arrangement will not be tested – but the four sites of the cluster will be sampled with the correct weight. In this process, one effectively generates a minimum spanning tree (with no loops) that connects every site on a cluster.

Note that this procedure checks (or grows) the cluster one growth shell at a time. Each growth shell is at a successive value of the minimum [50] or “chemical” [25] distance from the seed, so this method is useful for studying minimum distance problems including the fractal dimension  $d_{\min}$ . Like  $d_B$ , its value is not known exactly even in 2d and has to be determined by simulation. In 2d, its value is  $d_{\min} = 1.1307(4)$  [20].

### The Hoshen–Kopelman Algorithm

The Hoshen–Kopelman algorithm [29] has been a mainstay of work on percolation and is fairly well-known. It is described in some detail in the article of Stauffer in this work ► [Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation](#) and so will just be described briefly here.

The main idea of this algorithm is that a percolating system (say in two dimensions) can be examined or created a row at a time, and the cluster statistics can be updated just from the knowledge of the connections in the previous row. In  $d$ -dimensions, one must remember the state of the previous surface in  $(d - 1)$ -dimensions. The connections can be remembered in a look-up table, or a rooted-tree data structure can be used.

This algorithm can be used to analyze the clusters statistics of a given, fully populated system, or it can be

used as a very memory efficient scheme (especially in 2d) to generate and analyze on the fly a large system, since only the previous row (in  $2 - d$ ) needs to be remembered. Using this scheme, Tiggemann has simulated a lattice of  $4\,000\,000^2$  sites [73].

To actually identify all of the clusters on a given lattice, it is faster to sweep across the lattice using one of the neighbor-search algorithms above, because in the HK method the lattice would have to be swept a second time in order to get the most updated cluster labels.

Recently, Deng and Blöte [10] used a version of the Hoshen–Kopelman algorithm to determine  $p_c = 0.5927465(4)$  for site percolation on the square lattice, and  $p_c = 0.3116077(4)$  for site percolation on the simple cubic lattice, using a novel method of analysis based upon universal ratios of correlation functions and moment distributions. Tiggemann [73] has used a massively parallel version to study percolation on lattices in two, three, and four dimensions. A good discussion of the HK algorithm, its antecedents in the computer science field, and some of its modifications and extensions, is given by Martín-Herrero [41].

### Example: Critical Density Plots

As an example of an application of the above simulation method, we show in Fig. 1 the average density of clusters anchored to a single point and simultaneously to two points at the boundary of a square system. Here the density is defined as the number of times a given site is connected to the corresponding boundary point, divided by the total number of realizations. We use bond percolation with the growth algorithm and the FIFO (queue) method, with `xlist` and `ylist` large enough to remember the largest possible cluster. This allows us to transfer the coordinates of the cluster to the appropriate array (touching one anchor, touching the other anchor, or touching both anchors) after the cluster is grown. In this simulation, we don’t have to scan the entire lattice, but just use the two anchor points as seeds (`x0`, `y0`) for the cluster growth/identification.

To analyze this problem theoretically, it is convenient to put the system on the complex plane, and consider a half-infinite system  $y \geq 0$ , with the boundary placed on the real axis, and anchors at  $x_a$  and  $x_b$ . The results of this calculation can be transformed to a square by a standard conformal map  $w(z)$ , using the transformation of the density  $\rho(w) = (dw/dz)^{-h} (\overline{dw}/\overline{dz})^{-\bar{h}} \rho(z)$  where  $h = \bar{h} = (2 - D)/2 = 5/96$  for  $2 - d$  percolation, and  $D$  is the fractal dimension. Note that when the mesh size (lattice spacing) goes to zero with the boundaries fixed, the ac-



**Percolation Lattices, Efficient Simulation of Large, Figure 1**  
Simulation results for the average density of clusters touching the left anchor (top left), right anchor (bottom left), both anchors simultaneously (lower right), and prediction from Eq. (2) (upper right). From [35]

tual density of the clusters, being fractal, goes to zero. The density we consider is effectively renormalized to remain finite and non-zero in that limit.

For a single anchor placed at  $x_a$  on the real axis, the density at a point  $z = x + iy$  can be interpreted as the probability that the point  $z$  is connected to the point  $x_a$  – in other words, it is the two-point correlation function,  $P(z, x_a)$ . This quantity is predicted to be given by [35]

$$P(z, x_a) = \frac{c y^{11/48}}{|z - x_a|^{2/3}} \quad (1)$$

where  $y = (z - \bar{z})/(2i)$  and  $c$  is a non-universal, lattice-dependent constant. It turns out that this density is closely related to that of a dipole in electrostatics – it is  $y^{-5/48}$  multiplied by the potential of a dipole, raised to the  $1/3$  power.

Figure 1 shows the density contours in a square for single anchors  $P(z, x_a)$  and  $P(z, x_b)$ , and for two simultaneous anchors, which corresponds to the three-point correlation function  $P(z, x_a, x_b)$ . Now, it was observed that the three-point function is proportional to the square root of the product of the two-point correlation functions and the probability  $P(x_a, x_b)$  that  $x_a$  and  $x_b$  are connected together [35]:

$$P(z, x_a, x_b) = C \sqrt{P(x_a, x_b)P(z, x_a)P(z, x_b)}, \quad (2)$$

where  $C$  is a constant, valid as long as  $z$  is at least several lattice spacing from the anchor points  $x_a$  and  $x_b$ . Near the anchor points for a finite mesh,  $C$  is not constant but is a function of  $x_a, x_b$ , and  $z$ . When  $z$  approaches an anchor point, say  $x_a$ , it follows (for bond percolation) that  $P(x_a, x_a, x_b) = P(x_a, x_b)$ , and  $P(x_a, x_a) = 1$ , so at this point  $C$  is identically 1. However, when the mesh goes to zero,  $C$  is constant greater than 1 everywhere else.

Furthermore, rather surprisingly it was found that (away from the anchor points)  $C$  was the same for site and bond percolation, with the value  $= 1.030 \pm 0.001$ , and so appeared to be universal. After these numerical observations were made, it was shown theoretically that (2) indeed follows from conformal field theory using boundary operators [35], and the constant  $C = C_{222}$  is universal and given explicitly by [65]

$$\begin{aligned} C_{222} &= \sqrt{\frac{\Gamma(2/3)^3 \Gamma(5/3)^2}{\Gamma(1/3) \Gamma(4/3)^3}} \\ &= \frac{2^{7/2} \pi^{5/2}}{3^{3/4} \Gamma(1/3)^{3/2}} = 1.02992679 \dots \end{aligned} \quad (3)$$

### Excess Number of Clusters

Another example where precise simulations inspired a theoretical result is given by the problems of the excess number of clusters [92,93]. At the critical point, the number of clusters per lattice site  $n_c$  is a well-defined, finite, non-universal quantity. This quantity is known exactly for  $2 - d$  bond percolation on the triangular and square lattices; for the latter, Temperley and Lieb [72] found an integral expression for  $n_c$  which evaluates simply to  $(3\sqrt{3} - 5)/2 = 0.098076211 \dots$  [92]. Counting the critical clusters  $N_c$  on finite lattices of size  $W \times H$  with periodic boundary conditions, it was observed that

$$N_c = n_c HW + b(W/H) + \dots \quad (4)$$

where  $b(r)$  is a universal quantity that is a function of the aspect ratio  $r = W/H$  but independent of the underlying percolation type (in contrast to  $n_c$ , which is not universal and varies from system to system). The universality of  $b(r)$  was shown to follow as the singular part of the free energy [1], and  $b(r)$  has been calculated exactly for some geometries. For example, in the limit that  $r \gg 1$ , that is, for a cylinder,  $b(r)$  is proportional to  $r$  and is given simply by [34]

$$b(r) \sim \frac{5}{8\sqrt{3}} r. \quad (5)$$

This result is valid for all forms of critical  $2 - d$  percolation, is reminiscent of (but not identical to [32]) the num-



ber of wrapping critical clusters in a cylindrical system of length  $r \gg 1$ :

$$N_{\text{wrap}}(r) \sim \frac{1}{\sqrt{3}} r. \quad (6)$$

The universality of the excess cluster number also applies in higher dimensions, although no theoretical results are known there.

### Finding $p_c$ from the Leath Method

The Leath method can be used to generate single clusters in an empty lattice and find an unbiased measure of the size distribution. Say the procedure is started on a single seed. If the cluster grows to a size greater than or equal to some cutoff value  $s_{\text{max}}$ , it is stopped. The cutoff is chosen sufficiently small so that the growth will always stop before the boundaries of the lattice are reached. In that case, the statistics of the upper cumulative size distribution are unbiased by any boundary effects, and the only finite-size effects are those imposed by the cutoff value.

We define as usual

$$n_s(p) = \text{the number of clusters} \\ \text{containing } s \text{ sites, per lattice site} \quad (7)$$

then it follows that

$$P_s(p) = sn_s(p) \\ = \text{the probability that a given site} \\ \text{belongs to a cluster containing } s \text{ sites} \quad (8)$$

and

$$P_{\geq s}(p) = \sum_{s' \geq s} s' n_{s'}(p) \\ = \text{the prob. that a given site belongs to} \\ \text{a cluster containing } s \text{ or more sites.} \quad (9)$$

Here we are considering bond percolation; if it were site percolation, there would be an extra factor of  $p$  in  $P_s$  and  $P_{\geq s}$  reflecting the probability that the selected site is an occupied one. The quantity  $P_{\geq s}(p)$ , unbiased for all  $s < s_{\text{max}}$ , is determined directly by the growth of single clusters.

According to the usual scaling theory, in the scaling limit where  $s \rightarrow \infty$  and  $z = (p - p_c)s^\sigma$  is held constant,

$$n_s(p) \sim c_0 s^{-\tau} f(c_1(p - p_c)s^\sigma) \quad (10)$$

where  $c_0$  and  $c_1$  are the non-universal metric factors, while the exponents  $\tau$  and  $\sigma$ , and the scaling function  $f(z)$ , are universal. (This scale variable  $z$  should not be confused with the complex coordinate  $z$  above.) To define  $c_0$  and

$c_1$  uniquely, one can assume for example  $f(0) = 1$  and  $\int_{-\infty}^{\infty} f(z)dz = 1$ . It follows from (10) that

$$P_{\geq s}(p) \sim \frac{c_0}{\tau - 2} s^{2-\tau} g(c_1(p - p_c)s^\sigma) \quad (11)$$

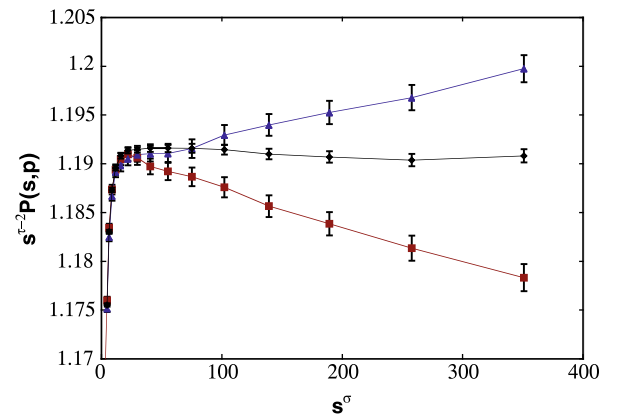
where  $g(z) = [(\tau - 2)/\sigma] z^{(\tau-2)/\sigma} \int_z^{\infty} \xi^{(2-\tau)/\sigma-1} f(\xi) d\xi$  for  $p > p_c$ , and similarly for  $p < p_c$ . Now it follows that if  $f(z)$  is analytic near  $z = 0$ , then  $g(z)$  is also, and we can carry out a Taylor-series expansion, yielding

$$P_{\geq s}(p) \sim s^{2-\tau} (A + B(p - p_c)s^\sigma + \dots) \quad (12)$$

where  $A$  and  $B$  are constants. Thus, a plot of  $s^{\tau-2} P_{\geq s}(p)$  vs.  $s^\sigma$  should give a straight line with a slope proportional to  $p - p_c$ .

### $p_c$ for the hcp and fcc Lattices

We illustrate this method for 3-d site percolation on the hcp and fcc lattices, from [40]. The lattice size was  $2048^3$ , created by using a virtual-memory scheme [89] that only assigns physical memory to cubes of the space as the clusters grows into them, which works well for the growth method because only a small fraction of the lattice is accessed by an individual cluster. The cutoff was  $s_{\text{max}} = 2^{21} = 2048576$ . The exponents  $\tau$  and  $\sigma$  are known exactly in two dimensions but not three, and by fitting the data to (12), those exponents can be found by this method as well. In Fig. 2, we show a plot of  $s^{\tau-2} P_{\geq s}(p)$  vs.  $s^\sigma$  for three close values of  $p$ , using  $\tau = 2.189$  and  $\sigma = 0.455$ . These exponent values were consistent with this data and also that of other 3-d systems, and compare with the val-



**Percolation Lattices, Efficient Simulation of Large, Figure 2**  
Single-cluster growth statistics for site percolation on the hcp lattice, with  $p = 0.1992600, 0.1992555$ , and  $0.1992500$  (top to bottom). Here  $P(s, p) \equiv P_{\geq s}(p)$ . From [40]



ues 2.18906(8) and 0.4522(9) respectively found by Balles-teros et al. [3], and 2.18958(9) and 0.4535(2) by Deng and Blöte [10]. The plot shows clearly that for large  $s$ , the behavior predicted by (12) is well followed. For  $s$  less than about 1000, there are significant deviations due to the finite-size effects of the lattice discreteness. These finite-size effects at  $p_c$  can be fit asymptotically by an equation of the form  $P_{\geq s}(p_c) \sim s^{\tau-2}(A + Cs^{-\Omega})$  with  $\Omega \approx 0.64$  [39].

First the simulations were run at the two outside values of  $p$ , and then (12) was used to extrapolate  $p_c = 0.1992555$ , which was then verified by a third run at this value, which is seen to be horizontal (for large  $s$ ) in the plot. Analyzing the errors yields  $p_c = 0.1992555(10)$ . A similar calculation for the closely related fcc lattice yields the slightly but statistically significant lower value  $p_c = 0.1992365(10)$ . For bond percolation, however, the thresholds for these two lattices at this level of precision are the same  $p_c \approx 0.120164$ , although they are likely to be different if measured to higher precision. Clearly, a very sensitive method, like the one presented here, is needed to distinguish such close thresholds.

Another application of single cluster growth has been to find amplitude ratios of the mean cluster size an equal distance below and above  $p_c$ , yielding a value  $\Gamma^-/\Gamma^+ = 163 \pm 2$  in  $2-d$  [33], much more precise than earlier determinations, and confirmed by high-order series expansions [33]. Finally, we note that a similar method of analysis has been applied to directed percolation in 2+1 dimensions [49] leading to the determination of the threshold and critical exponents to higher accuracy than previous works [22,74].

### Hull Walks and Hull-Generating Walks

The hull in two-dimensional percolation is the boundary between occupied and vacant “perimeter” sites of a percolation cluster. A typical cluster has both external and internal hulls, and an infinite cluster at the critical point has an infinite number of hulls within hulls. The fractal dimension of critical hulls was first conjectured (based upon simulations) to be simply  $7/4$  [60], and then this conjecture was proven theoretically first from field theory [59] and more recently using Stochastic Loewner Evolution (SLE) [67].

The “accessible” or Grossman–Aharony hull is the hull (not necessarily external) in which closed-off inlets or “fjords” are bridged and the hull shortened [23]. It turns out that this hull is also a fundamental measure of percolation clusters and at the critical point has a fractal dimension of  $4/3$ , identical to self-avoiding walks and the hulls of simple Brownian motion.

### Hull-Walk Algorithm

Hulls can be identified on an existing percolation cluster by carrying out a walk that follows the edges of the cluster, as first studied by Voss [76]. On the other hand, just like in the cluster growth algorithm, one can start with an unvisited (undetermined) lattice and decide upon the occupancy of the sites or bonds as they are encountered, and thus generate the hull at the same time as it is being identified. This idea was proposed for site percolation on the square lattice in [89], site percolation on the triangular lattice in [78], and bond percolation on the square lattice in [19].

For bond percolation, the easiest formulation of the walk is to follow a path that jumps between the centers of the occupied and vacant bonds on the perimeter – or, equivalently, between the bonds on the lattice and the dual lattice. For bond percolation on a square lattice, the paths also follow a square lattice, rotated at  $45^\circ$  from the bond lattice. When an occupied bond is encountered, the walk turns clockwise, while when a vacant bond (or bond on the dual lattice) is encountered, it turns counter-clockwise.

In the hull-generating procedure, when an UNVISITED bond is encountered, that bond is made OCC with probability  $p$  and the walk turns clockwise, and made VAC with probability  $1 - p$ , and the walk turns counter-clockwise. When an already visited site is encountered, the walk turns in such a way that the path always avoids itself. Following is a piece of a program that carries out these steps:

```
dir = 100; x = xo; y = yo;
do
{ x += dx[dir & 3];
  y += dy[dir & 3];
  switch (lat[x][y])
  { case UNVISITED:
    if (random() < probab)
    { lat[x][y] = OCC;
      ++nocc; --dir;
    } else
    { lat[x][y] = VAC;
      ++nvac; ++dir;
    } break;
    case VAC: ++dir; break;
    case OCC: --dir;
  }
} while ((x != xo) && (y != yo)
        && ((dir&3) != 0));
```

(P7)

Here we have rotated the original lattice by  $45^\circ$  so that the walk moves in horizontal and vertical directions and the bonds are effectively on the diagonals. The “& 3” construction is equivalent to “% 4” (modulo 4). The walk

ends when it returns to the starting point and it is going the same directions as it started out.

For lattices other than the square one, it is generally convenient to transform the lattice so that it fits on a square one. For example, a triangular lattice can be put on a square lattice with one diagonal bond put in. Then the hull walk has to be constructed between the centers of these bonds, with is rather intricate. An alternative approach is to remain on the simple square lattice with the walk moving in the four diagonal directions, but to make some bonds permanently occupied and/or vacant to simulate the particular lattices. For example, to create a triangular lattice, half of the horizontal bonds, alternating on each row, can be made permanently occupied. Making the same bonds permanently vacant gives the honeycomb lattice, etc.

For site percolation, one can use a variation of the above program in which the walk steps along occupied perimeter sites, always keeping the vacant sites on one side. Details are given in [89].

### Finding $p_c$ Directly from the Hull-Generating Walk

Starting from an UNVISITED lattice and a single seed, the hull-generating walk above will always close on itself, forming either external or internal hulls, depending upon the direction of closing. By making a simple hypothesis that at the critical point the internal and external hulls are equally likely (for large hulls), one can deduce an estimate for the critical point. For site percolation on a square lattice this method gave  $p_c = 0.59275(3)$  in work from over 20 years ago [82]. This approach has not been pursued further, nor have questions of its finite-size effects been explored.

For  $p$  away from  $p_c$ , the statistics of the internal vs. external hulls will be much different. It was found that the average number of occupied sites in a hull  $\langle s_H \rangle$ , for site percolation on the square lattice, satisfies [89]

$$\begin{aligned} \langle s_H \rangle_{\text{ext}}^- &\sim \langle s_H \rangle_{\text{int}}^+ \sim A|p - p_c|^{-2} \\ \langle s_H \rangle_{\text{int}}^- &\sim \langle s_H \rangle_{\text{ext}}^+ \sim B|p - p_c|^{-2} \end{aligned} \quad (13)$$

with  $A \approx 0.5$  and  $B \approx 0.004$ , where “int” and “ext” represent internal and external hulls, respectively, and + and – represent above and below  $p_c$ , respectively. (The simple exponent  $-2$  above follows from scaling relations of the hull exponents and  $D_H = 7/4$  [78,82].) These results imply that the average hull size shows an amplitude ratio of  $A/B \approx 125$ , reflecting the huge difference between the average size of these two kinds of hulls. While amplitude ratios play an important role in statistical mechanics [52], this ratio has not been studied further.

### The Enclosed Area Distribution

A more recent application of the hull-generating method has been to find the enclosed area distribution  $2 - d$ . (Even though both the cluster and the hull are fractal, with dimension  $91/48$  and  $7/4$ , respectively, the area enclosed by a hull is non-fractal and is proportional to the radius squared.) A simple argument from the scaling relation  $n_s \sim c_0 s^{-\tau}$ , the fractal relation  $s \sim A^{D/d}$  and the hyperscaling relation  $\tau - 1 = d/D$  implies that, in a critical system of total area  $A_{\text{tot}}$ , the number of clusters whose enclosed area is greater than  $A$  is given by [6]

$$N_{\geq A} \sim C A_{\text{tot}}/A \quad (14)$$

for  $A$  large compared to the mesh area  $A_0$  and small compared to  $A_{\text{tot}}$ . (For  $d > 2$ ,  $A$  represents the volume of an enclosing sphere or rectangular solid, and  $C$  is different.) The area distribution can be found from an ensemble of individual closed hulls (loops) generated by the hull-walk algorithm, and it is a simple addition to the program to calculate the enclosed area of the walk on the fly, while the walk is carried out. The coefficient  $C$  was found from simulations to be a universal constant  $0.022976(5)$ , and proven theoretically to be simply [6]

$$C = 1/(8\sqrt{3}\pi) = 0.022972037 \dots \quad (15)$$

via a conformal transformation to the problem of the number of clusters wrapping a cylinder given above in (6). Equation (14) represents a completely universal formulation of the size distribution at criticality (in contrast to  $n_s \sim c_0 s^\tau$  which involves both the non-universal  $c_0$ , and the non-universal measure of the size,  $s$ ). Another way to express (14) is in Zipf’s-law form: if you rank-order all the hulls in the system by their enclosed area, then the area of the  $n$ th ranked hull is inversely proportional to  $n$  and is given by  $C A_{\text{tot}}/n$  for large  $A$ . Other forms of the universal size distribution are given in [93].

### Applications of the Hull-Generating Walk to Crossing Problems

Hull generating walks can be used to efficiently test for crossing or spanning. For example, consider a rectangular system. A walk is started in the lower left-hand side, and represents the boundary between the occupied bonds above it and the vacant bonds below it. If the walk reaches the right boundary before reaching the top boundary, then there is horizontal crossing, while if it reaches the top before reaching the right-hand side, then there is no horizontal crossing. This process is more efficient than filling the entire lattice with clusters, because only the bonds along

the hull are simulated. It allowed a sensitive test [83,86] of the finite-size corrections to the crossing probability for a rectangle, which in an important development in percolation theory was found by Cardy to be given by [5]:

$$\Pi_h(r, p_c) = C_{112}^2 \lambda^{1/3} {}_2F_1(1/3, 2/3; 4/3; \lambda), \quad (16)$$

with  $C_{112}^2 = 2\pi\sqrt{3}/\Gamma(\frac{1}{3})^3 = 0.56604668\dots$ , where  ${}_2F_1$  is the hypergeometric function and the subscript  $h$  signifies the horizontal crossing probability. Here  $r$  is the aspect ratio of the rectangle, and is related to  $\lambda$  by  $\lambda = ((1-k)/(1+k))^2$  and  $r = 2K(k^2)/K(1-k^2)$ , where  $K(m)$  is the elliptic integral. This parametrization can be simplified to  $r = K(1-\lambda)/K(\lambda)$  and inverted explicitly as [85]  $\lambda = \vartheta_2^4(\exp(-\pi r))/\vartheta_3^4(\exp(-\pi r))$  where  $\vartheta_n(q)$  are elliptic theta functions. Finally, the parameter  $\lambda$ , which represents the cross-ratio of the coordinates of the corners of the rectangle when mapped to the half plane, can be eliminated to yield a closed-form explicit expression for  $\Pi_h(r, p_c)$ , differentiated with respect to  $r$  [85]:

$$\begin{aligned} \frac{\partial \Pi_h(r, p_c)}{\partial r} &= -\frac{1}{3} \pi C_{112}^2 \vartheta_1'(e^{-\pi r})^{4/3} \\ &= -\frac{2^{4/3}}{3} \pi C_{112}^2 \eta(ir)^4 \end{aligned} \quad (17)$$

where  $\eta(\tau)$  is the Dedekind eta function. Equation (17) implies the series

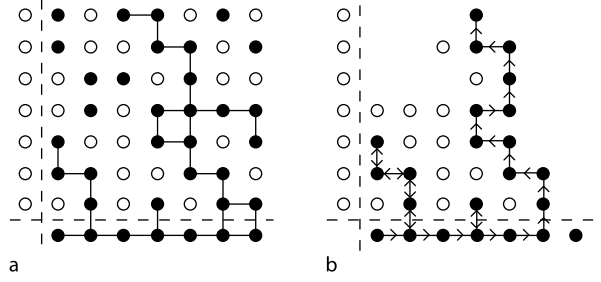
$$\Pi_h(r, p_c) = \frac{2^{4/3} \pi C_{112}^2}{3} \left( e^{-\pi r/3} - \frac{4}{7} e^{-7\pi r/3} \dots \right). \quad (18)$$

The prediction for  $\Pi_h(1, p_c) = 1/2$  the a system with a square boundary for site percolation on a square lattice of size  $L \times L$  was verified in [83]. The hull-walk used in that work is illustrated in Fig. 3, here rotated so that crossing is considered in the vertical rather than horizontal direction. The behavior that crossing translates into the walk hitting one boundary before the other is exactly analogous to problems solved in SLE.

This work showed that the finite-size corrections to  $\Pi_h(1, p_c) = 1/2$  are for large systems described by

$$\Pi_h(1, p_c) = 1/2 + 0.319/L + \dots \quad (19)$$

The numerical results were not consistent with a significant contribution from the “irrelevant” scaling variable  $L^{-0.85}$  [84] and later it was shown that indeed because of the symmetry of the square system, the irrelevant term does not contribute here [30]. A consequence of (19) is that several estimates of  $p_c$  based upon measurements of  $\Pi_h(1, p_c)$  do not converge with the usual scaling  $\sim L^{1/\nu}$  but instead with the scaling  $L^{-1-1/\nu}$  [83].



**Percolation Lattices, Efficient Simulation of Large, Figure 3**

**Hull walk for a test for vertical crossing for site percolation on the square lattice. a** all clusters connecting the bottom to the top. **b** the equivalent hull that would be generated from the site-percolation hull-generating walk with the same probability. *Solid circles: occupied sites. Open circles: vacant sites.* From [83]

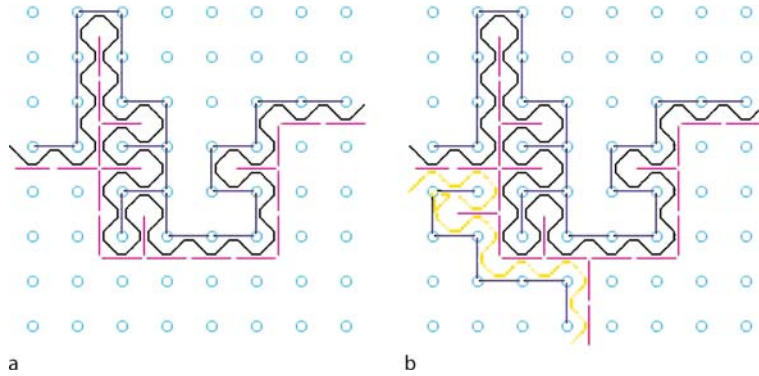
The hull method also is very efficient for exact enumeration. Basically, by stepping through every possible hull, one can determine the polynomials for  $\Pi_h(r, p)$  as a function of  $p$ , for a fixed  $r$  and system size. This way  $\Pi_h(1, p)$  could be found for square system size  $L \times L$  for  $L$  up to 7, [83,94] which would be very difficult to do with a complete exact enumeration, as this would require  $2^{49}$  realizations of the lattice. For the first few values of  $L$ ,  $\Pi_h^{(L)}(1, p)$ , written as a series in  $p^i q^{L^2-i}$  where  $q = 1 - p$ , is given by

$$\Pi_h^{(2)}(1, p) = 2p^2 q^2 + 4p^3 q + p^4 \quad (20)$$

$$\begin{aligned} \Pi_h^{(3)}(1, p) &= 3p^3 q^6 + 22p^4 q^5 + 59p^5 q^4 \\ &\quad + 67p^6 q^3 + 36p^7 q^2 + 9p^8 q + p^9. \end{aligned} \quad (21)$$

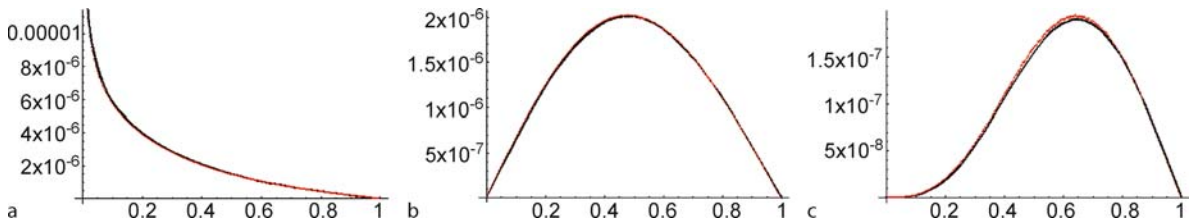
Polynomials for  $L$  up to 4 are given in [56] and those for  $L$  up to 7 can be found in [94]. These results can be used for a variety of studies; for example, looking at the kurtosis of the distribution, it was shown [84] that the distribution of first crossings  $(\partial/\partial p)\Pi_h(r, p)$  is not a gaussian curve as had been previously thought. This work was followed by a more general mathematical proof [4] that the tails behave like  $\ln[(\partial/\partial p)\Pi_h(r, p)] \sim -L/\xi \sim -L|p - p_c|^{4/3}$  where  $\xi \sim |p - p_c|^{-\nu}$  with  $\nu = 4/3$  in 2d is the correlation length.

In recent work [66], the hull-generating method was used to test a generalization of Cardy’s formula that describes the probability density that crossing clusters have lower edges at  $y = a$  and  $y = b$  on the left and right-hand boundaries, respectively, with various conditions on whether the cluster touches the bottom. Figure 4 shows how this problem is simulated by a hull walk, here for bond percolation with the bonds themselves horizontal



Percolation Lattices, Efficient Simulation of Large, Figure 4

**a** The hull-generating walk (in black) used to test for a cluster, whose lower edge is half-way up the left-hand side, crossing to the right-hand side, and to find the distribution of values  $y$  for where it hits on the right-hand side. **b** an additional walk (in yellow) to check that there are no other clusters crossing below the given walk. From [66]



Percolation Lattices, Efficient Simulation of Large, Figure 5

Measurements (points) and theory (lines) for the distribution of the lower boundary on the right-hand side, of clusters whose lower boundary on the left-hand side is at  $y = 1/2$ . **a** Clusters that touch the bottom, **b** no restriction on the crossing of the clusters, and **c** clusters that cross from left to right but do not touch the bottom, and have no crossing clusters below them. From [66]

and vertical, and the steps along the diagonals. Figure 5 shows excellent agreement between the simulations and the theory. The hull-generating walk proved very efficient for this problem, since walks that hit one of the forbidden boundaries (used to enforce the crossing criteria) were stopped without generating the rest of the hull. Thus, in a few day's of computer work, it was possible to simulate  $3.3 \cdot 10^{11}$  hulls on a lattice of  $512 \times 512$  bonds, something that would be nearly impossible to carry out if all lattice bonds were considered in the simulation.

### Gradient Percolation

Sapoval, Rosso and Gouyet [60] first considered percolation in a gradient, and showed that in 2d it is an efficient way to find that percolation threshold [57]. In this approach, a rectangular system is set up with a linear gradient in  $p$ , going from 0 to 1 as  $y$  goes from 0 to 1. There will be a percolating cluster connected to the upper boundary, and the hull of that cluster will sample values of  $p$  that are close to  $p_c$ . Sapoval et al. [60] showed that the hull will stay within a relatively small region  $L^{4/7}$  of the lattice of inverse

gradient  $|\nabla p|^{-1} = L$ . Therefore, as  $L \rightarrow \infty$ , the “frontier” will be localized about  $p_c$ .

For a finite  $L$ , two measures of  $p_c(L)$  are (i), the average value of  $p$  of all the occupied+vacant sites (or bonds) of the hull, or, (ii) just the fraction of occupied to total bonds along the hull:

$$p_c^{\text{est}}(L) = \frac{n_{\text{occ}}}{n_{\text{occ}} + n_{\text{vac}}} . \quad (22)$$

For large systems (small gradients), these measures should be asymptotically equivalent. Simulating system of size up  $1000 \times 1000$  for site percolation on the square lattice, Rosso et al. found that the estimates fell on a straight line when plotted as a function of the gradient  $1/L$ , with as extrapolated value of 0.592805(10), slightly higher than the values we have seen above.

By combining gradient percolation with the hull-generating walk, one can create a very efficient and simple method to determine percolation thresholds in two dimensions [90]. Basically, the program (P7) is used, with `prob` now a function of  $y$ , and periodic boundary conditions in the horizontal direction. The “front” ( $x$  coor-

dinate) of the walk is kept track of, and when it reaches a new value, the column (all values of  $y$ ) with the value of  $x$  mapped on to the periodic lattice is cleared out and returned to the UNVISITED state. This is allowed because the walk snakes back a maximum distance of the order of its width in the  $y$  direction, so that sites or bonds behind that can be forgotten. Thus the simulation runs continuously, effectively simulating an infinitely wide system.

To start the walk, one has to make a vertical column on the left of all occupied sites or bonds above the starting point, and vacant ones below it. This will keep the walk from closing on itself at the beginning. Data from the early part of the simulation can be thrown away to eliminate any bias that it causes.

This method was used to find precise thresholds, some up to seven significant figures, for a variety of two-dimensional lattices, including the Archimedean lattices for site percolation [69], and the kagomé lattice for bond percolation [91]. These results are useful for understanding how thresholds depend upon the lattice structure, and to test conjectures for the values of the thresholds (see [62]).

The question of the convergence of the estimates is open: for many systems, the convergence behavior seems to change from  $1/L$  for smaller  $L$  to a different behavior (or perhaps the same  $1/L$  behavior but with a different coefficient) for larger  $L$ . In fact, for site percolation on the square lattice, it turns out that the linear behavior seen by Rosso et al. breaks down for  $L$  larger than about 1000, and the curve levels off, extrapolating to a value  $\approx 0.5927465$ , close found by other methods. The understanding of the convergence of this method remains an open problem.

The errors can be determined easily by looking at batches of results, and are proportional to  $(n_{\text{occ}} + n_{\text{vac}})^{-1/2}$ . The proportionality constant is of order 1, indicating a very efficient method, and grows slowly with increasing  $L$ , implying that with increasing  $L$ , somewhat more work is needed to achieve the same level of precision.

### Example: the Critical Surface for the Checkerboard Lattice

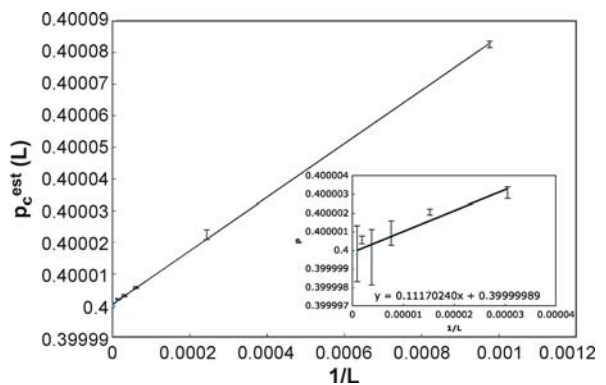
As an example of this method, we consider the checkerboard lattice, that is a square lattice with four different probabilities  $p_1, p_2, p_3$ , and  $p_4$  around each colored square. According to a conjecture by Wu concerning the more general  $q$ -state Potts model, here specialized for  $q = 1$ , the critical surface satisfies the formula [80]

$$1 - (p_1 p_2 + p_1 p_3 + p_1 p_4 + p_2 p_3 + p_2 p_4 + p_3 p_4) + p_1 p_2 p_3 + p_1 p_2 p_4 + p_1 p_3 p_4 + p_2 p_3 p_4 = 0. \quad (23)$$

This result does not appear to follow directly from duality, in contrast to all other exact results known in percolation [70,79,95]. However, it reduces to the known exact results for the inhomogeneous honeycomb lattice (letting  $p_4 = 0$ ), the inhomogeneous triangular lattice (letting  $p_4 = 1$ ), and the dual checkerboard lattice ( $p_2 = 1 - p_1$ ), and duality in the sense that  $p_i \rightarrow 1 - p_i$  also satisfies this formula. It is the most general relation of this form, linear in all of the individual probabilities, that satisfies these requirements, but there is no obvious reason why it has to be linear in this way.

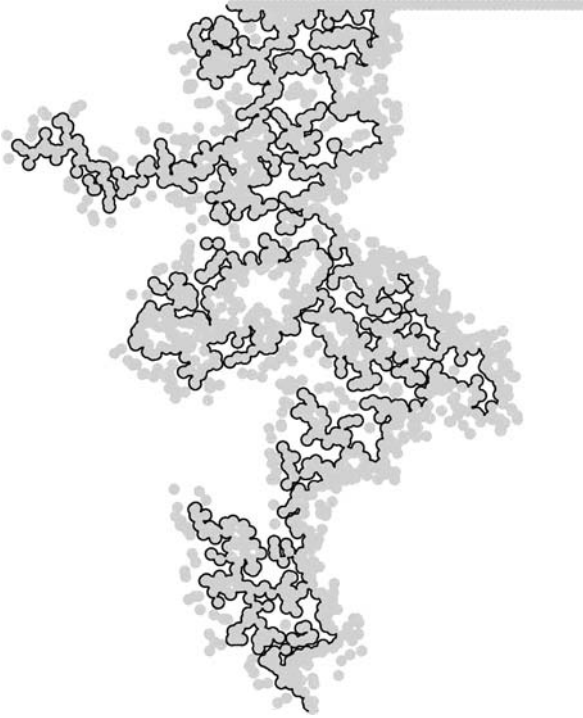
It seems that this result has not been tested numerically in the past. Here we investigate one case  $p_1 = 73/90 = 0.811111\dots$ , and  $p_2 = p_3 = p_4 = p$  using the hull-gradient method. According to the conjecture (23),  $p$  should equal 0.4 exactly.

In order to test this prediction, we fix the value of  $p_1$  on every fourth bond, while for the rest of the bonds we allow  $p$  to follow the gradient, and use the  $n_{\text{occ}}$  and  $n_{\text{vac}}$  for these bonds to estimate the critical value of  $p$  by (22). Figure 6 shows the plot of  $p_c^{\text{est}}(L)$  that follows for different values of the inverse gradient  $L$ . The lattice was  $16384 \times 16384$ , but we were able to go to inverse gradients as large as  $L = 524288$  without having  $y_{\text{max}} - y_{\text{min}}$  or  $x_{\text{front}} - x$  exceed 16384. We also used a periodic scheme in the vertical direction in order to have the walk automatically adjust to its own position. In this particular case, the linear dependence of  $p_c^{\text{est}}$  upon the gradient  $1/L$  seems to hold, with an extrapolated value of  $0.39999989(20)$ . To achieve these very small error bars, a total of  $n_{\text{occ}} + n_{\text{vac}} = 10^{14}$  random numbers were generated, one for each time the walk encountered an unvisited bond whose state was not previously determined. Thus, Wu's conjecture is numeri-



Percolation Lattices, Efficient Simulation of Large, Figure 6  
Results of gradient percolation study for the checkerboard lattice with  $p_1 = 73/90$ , and  $p_2 = p_3 = p_4$  predicted to be 0.4 by (23). From [63]





**Percolation Lattices, Efficient Simulation of Large, Figure 7**  
Frontier (hull) of the percolating region for the continuum percolation of overlapping disks in a gradient in the horizontal direction. From [53]

cally confirmed to high accuracy for this point. Additional points are tested in [63].

The hull-gradient method has been generalized to continuum percolation, illustrated in Fig. 7, yielding the most precise known value for the fractional critical coverage  $\phi_c = 0.6763475(6)$  [53,54].

Note, when doing work like this, it is imperative to use a high-quality random number generator, and not ones typically incorporated in computer languages or compilers. For much of our own work reviewed here, we used a four-tap shift-register sequence random number generator based upon the exclusive-or operation, with maximum lag of 9689 and cycle  $2^{9689} - 1$  [87].

### Simulating the Grossman–Aharony Accessible Hull

It does not seem possible to make a random walk process that generates the accessible hull of a percolation cluster (the hull in which all “fjords” are closed off) directly, because of the long range correlations. However, it is possible to generate this hull by carrying out two walks: the first to generate say the outside hull of a cluster, and then a second walk that encircles the first, and can jump across the

fjords. In this way, samples of a walk that are equivalent to the two-dimensional self-avoiding walk can be generated easily.

In the same way, a second walk can be added to gradient percolation (delayed behind the first walk by the correlation length), and the second walk will trace out the accessible hull of the frontier in the gradient. In this, way, an infinitely long accessible hull can be made (essentially one dimensional, however, because of the effect of the gradient).

### The Microcanonical-Canonical Method

Here we discuss the method of Newman and Ziff [44] to simulate percolation that, unlike other methods considered so far, allows one to simulate problems for all values of  $p$  through one simulation.

The idea is to start with an empty lattice, and add one site or bond at a time, and update the cluster connectivity on the fly, somewhat like the Hoshen–Kopelman method, but applied to clusters rather than rows. The quantity of interest is stored as a function of  $n$ , the number of added sites or bonds. Call this quantity  $Q_n$ , which represents the fixed- $n$ , or “microcanonical” value of  $Q$ . Then, for a given probability, the “canonical”  $Q(p)$  follows from convolving with the binomial distribution

$$Q(p) = \sum_{n=0}^N \binom{N}{n} p^n (1-p)^{N-n} Q_n, \quad (24)$$

where  $N$  is the total possible number of sites or bonds in the system. Note that these approaches have also been described as “canonical” and “grand-canonical”, by considering  $s$  as representing the number of particles in the system [64]. Here, we are thinking in energetic terms, along the lines of the Potts model representation of percolation.

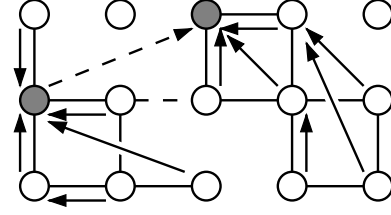
An example of this has already effectively been given above in the  $p^n q^{N-n}$  series of  $\Pi_h^{(L)}(1, p)$  (20) and (21), with  $N = L^2$ . Consider the case  $L = 2$ , and let  $Q(p) = \Pi_h^{(2)}(1, p)$ . The coefficients in (20) are precisely the number of ways of having horizontal crossing with  $n$  occupied sites and  $N - n$  vacant sites. Because there are  $\binom{N}{n}$  possibilities of having  $n$  sites in the system, it follows that the  $Q_n$  are just these coefficients divided by  $\binom{N}{n}$ . Therefore, for this system  $Q_0 = Q_1 = 0$ ,  $Q_2 = 1/3$ , and  $Q_3 = Q_4 = 1$  (the latter reflecting the fact that with three or four occupied sites, there will always be crossing). Then, the convolution in (24) is formally identical to  $\Pi_h^{(2)}(1, p)$  given in (20).

The antecedents of this method in the literature are many. The idea of extrapolating results of simulations

to different values of  $p$  is reminiscent of the histogram method [31]. The representation of clusters as a tree structure and some of the update bookkeeping steps are reminiscent of the Hoshen–Kopelman method [29]. Similar tree structure representations of clusters have been used in kinetic gelation models [8]. Finally, the idea of adding one occupied site or bond at a time was suggested in a problem in [16]. But [44] seems to be the first place that all these ideas were put together along with the convolution (24) and used to find results for some quantity for all values of  $p$ . These ideas been incorporated in an extensive discussion of this method in are Gould, Tobochnik, and Christian [17].

For definiteness we consider bond percolation. Initially, no bonds exist, and all sites are clusters of size one and each has a different label. (Again, the size is the number of sites the cluster contains). When a new bond is added, it can either connect sites belonging to the same cluster, in which case nothing needs to be done, or it can connect sites from two different clusters. In the latter case, these two clusters are combined into one by a union operation. For efficiency, the smaller cluster is incorporated into the larger one.

A simple approach to carry out this operation is to label each site of the lattice by an index representing the cluster it belongs to, and having a look-up table that registers the number of sites in each cluster. When a new bond connects sites of two different indices, the look-up table tells which of the two is smaller, and then a neighbor search like (P2) or (P3) can be used to relabel all sites of the small



**Percolation Lattices, Efficient Simulation of Large, Figure 9**

Tree data structure, shown the merging off the cluster of six sites on the left with the cluster of seven sites on the right, due to the addition of the new bond (dashed bond). Arrows show the directions of the links. From [44]

cluster to the index of the larger cluster. The appropriate updates to the look-up table are then made.

However, this method is somewhat slow because a given site is relabeled several times, and it can be improved by having a cluster remembered as a tree structure and linking the root of the smaller cluster to that of the larger one, as shown in Fig. 9.

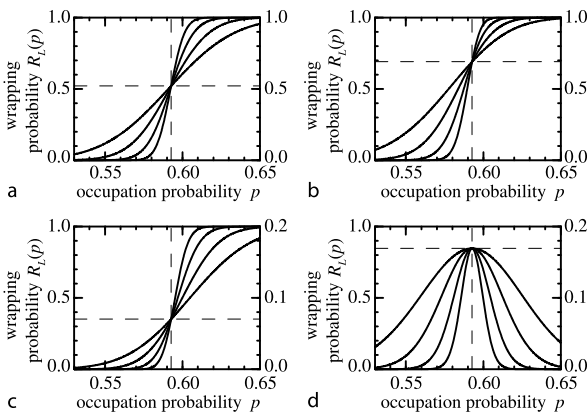
Here we describe a recursive program to carry out the “union-find” operation for bond percolation. More details are given in [45], which discusses the algorithm for site percolation.

First of all, in contrast to the programs given above, here we use a one-dimensional array `ptr[r]` to represent a system of any dimensions. An array `nn[i][dir]` is constructed ahead of time that tells the nearest-neighbors of every point  $i$  in the system, and thus can be set up for any boundary condition. This array is only used to decide which pair of sites a given bond connects. Sites are indexed with a single signed integer label for speed, taking values from 0 to  $N - 1$ .

The array `ptr[]` serves triple duty: for non-root occupied sites it contains the label for the site’s parent in the tree (the “pointer”); root sites are recognized by a negative value of `ptr[]`, and that value is equal to minus the size of the cluster; for unoccupied sites `ptr[]` takes the value `EMPTY`, which is defined as some value such as  $-N - 1$  that is never reached by any of the roots.

We define a function which performs the “find” operation, returning the label of the root site of a cluster, as well as accomplish path compression. The version we use is recursive:

```
int findroot(int i)
{
    if (ptr[i] < 0) return i;
    return ptr[i] = findroot(ptr[i]);
}
```



**Percolation Lattices, Efficient Simulation of Large, Figure 8**

Wrapping probabilities  $R_L(p) \equiv \Pi_{\text{wrap}}(p_c)$  calculated using the Newman–Ziff algorithm, for  $L \times L$  tori with  $L = 32, 64, 128$ , and  $256$ , for wrapping (a) along a specified axis, (b) along either axis, (c) along both axes, and (d) along one axis but not the other. The dotted lines denote the expected values of  $p_c$  and  $\Pi_{\text{wrap}}(p_c)$ . The curves are sharper as  $L$  increases. From [44]

When the recursion is “unwound”, all the sites of the links are relabeled to point to the new root. This seems to result in an optimal amount of relabeling to make this process run quickly.

The code to perform the actual algorithm is quite brief. Ahead of time, an ordered list of all the bonds is made. A given new bond connects the two neighboring sites `s1` and `s2`. The function `findroot()` is called to find the roots of each of the two sites. If amalgamation is needed, it is performed in a weighted fashion, smaller clusters being added to larger (bearing in mind that the value of `ptr[]` for the root nodes is *minus* the size of the corresponding cluster). Following is the main code to accomplish that:

```
r1 = findroot(s1);
r2 = findroot(s2);
if (r2!=r1)
    if (ptr[r1] > ptr[r2]) {
        ptr[r2] += ptr[r1];
        ptr[r1] = r2;
    } else {
        ptr[r1] += ptr[r2];
        ptr[r2] = r1;
    }
```

There are also easy techniques to check for crossing or wrapping during this process. One stores the number of such events as a function of  $n$ , the number of bonds put down, and then uses the convolution (24) to find the desired quantity as a function of  $p$ . Of course, the procedure must be repeated many times to get good statistics for all  $n$ . The result of a test of wrapping around a torus is given in Fig. 8.

The main point of this algorithm is that it can find the  $Q_n$  in a time that is very nearly linear in the number of lattice sites. Once the  $Q_n$  are found and stored in an appropriate array, they can be used to find various properties of the system for any  $p$ .

### Other Numerical Techniques

In this section we briefly mention some other numerical techniques that have been applied to percolation.

#### The Binary Search Method

In this method, a random number  $p_i$  uniform in  $(0, 1)$  is assigned to each site (say) of the lattice. One chooses a value of  $p$  such that sites with  $p_i < p$  are assumed to be occupied, and checks for percolation. By making a binary search up and down in  $p$  (with the same  $p_i$  assigned to each site), in about 20 steps the probability where that particular sample first percolates can be found to six significant

digits. Repeating this for many samples and averaging the results yields the *average* estimate for  $p_c$ .

If horizontal crossing is used for the criterion for percolation, then this average corresponds to  $\int_0^\infty p(\partial/\partial p)\Pi_h(p, r)dr$ . When systems of different sizes are simulated, finite-size scaling of this quantity can be used to extrapolate the estimate to  $L \rightarrow \infty$ . In general, finite-size scaling implies that estimates converge to  $p_c$  as  $L^{-1/\nu}$  [68], but for certain symmetric systems, such as a square boundary for site percolation on a square lattice, the convergence goes as  $L^{-1/\nu-1}$  [94] and even faster for wrapping around a periodic system [45].

This method is quite efficient in finding  $p_c$  since for a lattice of  $N = L^2$  sites, the total amount of time to measure one sample grows only as  $N \ln N$ . It has been used in numerous studies of percolation in various dimensions [68].

### Lattice-Less Methods

Vollmayr [75] introduced the idea that by using a kind of random number generator (effectively a very non-linear function) that takes as an input the coordinates of a site, and outputs a uniform random number that has no correlations with the random numbers that results from one of the neighbors, one does not have to remember the occupancy state of a given site. However, for most problems, one does have to remember whether that site has been visited (or checked as in the cluster algorithms), so techniques to remember the latter have to be used, some of which are discussed in [48]. There are several problems where one does not need to remember which sites have been visited, so for these problems this technique is particularly memory efficient. One example is the problem of finding the enclosed area of a hull-generating walk: here it is not necessary to remember if a site have been visited or not, and the method can be used to simulate any size system. Another example is just finding the end-to-end length of a very long walk, for a fractal measurement. One is only limited by the computational time available.

A drawback of this method is that it requires this special type of random number generator. Vollmayr uses a generator related to data encryption, which evidently has the necessary properties and is sufficiently uncorrelated and fast (though not as fast as typical random number generators used for Monte Carlo simulations). However, more work needs to be done to study this generator's quality for these types of problems.

Osterkamp, Stauffer and Aharony [46] introduced a related idea, using the feature of congruential random number generators that one can jump ahead or behind any

number of steps in the random number sequence by making an appropriate modification to the multiplier. Using this they were able to simulate diffusion on percolation clusters on high-dimensional (virtual) lattices as large as  $420^7 \approx 2.3 \cdot 10^{18}$  sites. This is another problem for which there is no need to remember which site has been visited before, so no list is needed, and the program required very little memory.

### Conductivity and Backbones

An important application of percolation is to flow and resistance problems. The conductivity of a percolating network (say with occupied bonds replaced by identical resistors) goes to zero as the percolation threshold is approached from above, and much work has been done in studying that process. The simulations are based upon solving Kirchhoff's equations around each vertex of the lattice.

When considering conductivity of a percolation cluster, the role of different bonds becomes evident. One can define conductivity between two points far apart in a cluster, or alternatively between two opposite edges of a bounded system (the “bus-bar” problem). In either case there will be a backbone that carries the current, and dangling ends that are only singly connected to the backbone, which can be removed. Within the backbone, bonds can be classified into different categories, with the “red” bonds being ones the “hottest” in that cutting any one of them will break the flow of current [50].

To find the backbone, several methods have been used, including Tarjan's method [71], burning algorithms [27], and matching algorithms [43]. Grassberger has introduced an efficient hull-walk based algorithm which however works only in two dimensions [20]. Once the backbone is found, the conductivity can be estimated efficiently in 2d by the algorithm of Lobb and Frank [38], which reduces the lattice by successive use of a star-triangle transformation, or in general by finite-element methods to solve the Kirchhoff equations. Conductivity can also be studied by considering the properties of random walks on the percolation cluster [28].

In 2d at the critical point, the backbone has a fractal dimension  $D_b = 1.6432(8)$  [21] or  $1.6434(2)$  [9] such that in a system of length scale  $L$ , the number of backbone bonds grows as  $L^{D_b}$ . The red sites scale simply as the inverse of the correlation-length exponent,  $D_{\text{red}} = 1/\nu$  [7], which equals  $3/4$  in 2d. The conductivity at the critical point scales with the system size as  $L^t$  with  $t/\nu = 0.9826(8)$  [21]. At one time there was a great deal of interest in studying this exponent because of the Alexander–Orbach con-

jecture [2] that  $t/\nu$  was equal to 1, which however was proven (numerically) to be incorrect in five back-to-back papers in Physical Review B in 1984 [26,28,38,55,81].

### Conclusions

This article describes a number of algorithms and programming techniques to study cluster statistics, crossing problems, area distributions, etc. of percolation. By no means did it cover all of them; having been a very active field of research for 50 years, there are many other methods and techniques that have been proposed and studied. Some applications of numerical techniques are also presented.

An example of the development of the substantial numerical work done in this field is provided by the determination of the threshold for site percolation on the square lattice, whose value has not been derived theoretically and must be found by simulation. After starting out at 0.581(15) in 1961 in the first Monte-Carlo determination [14], and after dozens of advances in the following three decades, by the early 1990s the six-digit value 0.592746 was achieved [83]. Yet, in the 16 years since then, while that value has been confirmed, the seventh digit still has not been agreed upon – the various determinations, many quoted above, fall in the range 0.5927460–0.5927466. Although one might think that with all the advances in computer power and algorithms that have occurred over these years, it would have been fairly easy to extend this result further, it turns out to be harder than might have been anticipated, because of uncertainties in the finite-size corrections and in the quality of random-number generators, which seem to be significant at this level of precision. Such high precision values are in fact necessary for precise studies of critical behavior, where simulations involving  $10^{13}$ – $10^{14}$  sites are not uncommon, and in 2d, site percolation on the simple square lattice remains one of the most popular models for various reasons, despite of the fact that exact thresholds are known for other  $2 - d$  system.

### Future Directions

Work in this field remains quite active and there are many interesting questions that are still unanswered. Convergence of many of the estimates, precise values of sub-leading (including irrelevant) exponents, more accurate calculation of the scaling functions and amplitude ratios are some such questions. For percolation thresholds, continued study of thresholds can perhaps lead to new exact results and in any case can help advance the understanding of why particular lattices have the thresholds that they



do. The combination of the efficient techniques that have been developed and improved upon over the years and the availability of powerful computers should allow many of these questions to be investigated fairly easily today. Here is a sampling of some specific questions for future study based upon the work discussed above:

- In the Newman–Ziff study of percolation around tori, it was found that the wrapping probabilities (for various situations, such as “either” or “one-way”) approach their theoretical values [51] as  $L^{-2}$ , implying that the estimates for  $p_c$  converge to the actual value as  $L^{-11/4}$ . However, no theoretical justification was found for this result. Note that additional numerical work showing fair agreement with this scaling for a variety of lattices was done recently by Parviainen [47].
- As mentioned above, the estimate of  $p_c$  for the hull-gradient method seems to converge as the reciprocal of the gradient for many systems (as in Fig. 6), while in others it changes its behavior and even is non-monotonic. Additional precise measurements, including tests with different random number generators, on a variety of systems can help elucidate this question. Another interesting question is the effect of the angle of the gradient with respect to the axis for various lattices. Some work along these lines for the kagomé lattice was done in [69].
- When written in terms of the enclosed area distribution, the size distribution follows the Zipf’s-law form (15) above which is an entirely universal form with no metric factors. Only preliminary studies have been made on this quantity away from  $p_c$ , and a simple exponential scaling curve is possibly seen [88]. Clarification of this behavior (and in relation to the Kunz–Souillard [36] form of the percolation scaling function which predicts exponential behavior in the size, not the area) is needed. Note there are other measures of the area that can be used, such as that of the enclosing Grossman–Aharony hull, that would also be interesting to study.

Recent interest in percolation by mathematicians, following the developments in Stochastic Loewner Evolution, will undoubtedly lead to many more simulation studies and new algorithms in percolation, such as [15] which concerns geometry of clusters on the closely related Potts model. It is interesting to note that in the percolation case, SLE develops a theory for the continuum limit of precisely the hull walks which were first introduced as a computational technique in this field more than two decades ago [19,78,89].

## Acknowledgments

This work was supported in part by the National Science Foundation under grant no. DMS-0553487. Comments by D. Stauffer and P. Kleban were highly appreciated.

## Bibliography

### Primary Literature

1. Aharony A, Stauffer D (1997) Test of universal finite-size scaling in two-dimensional site percolation. *J Phys A* 30:L301–L306
2. Alexander S, Orbach R (1982) Density of states on fractals: fractons. *J Phys Lett* 43:L623
3. Ballesteros HG, Fernández LA, Martín-Mayor V, Muñoz Sudupe A, Parisi G, Ruiz-Lorenzo JJ (1999) Scaling corrections: site percolation and Ising model in three dimensions. *J Phys A* 32: 1–13
4. Berlyand L, Wehr J (1995) The probability distribution of the percolation threshold in a large system. *J Phys A* 28:7127–7133
5. Cardy JL (1992) Critical percolation in finite geometries. *J Phys A* 25:L201–206
6. Cardy J, Ziff RM (2003) Exact area distribution for critical percolation, Ising and Potts model clusters. *J Stat Phys* 110:1–33
7. Coniglio A (1982) Cluster structure near the percolation threshold. *J Phys A* 15:3829–3844
8. de Freitas JE, Lucena LS (2000) Equivalence between the FLR time-dependent percolation model and the Newman–Ziff algorithm. *Int J Mod Phys C* 8:1581–1584
9. Deng Y, Blöte HWJ, Nienhuis B (2004) Backbone exponents of the two-dimensional  $q$ -state Potts model: A Monte Carlo investigation. *Phys Rev E* 69:026114
10. Deng Y, Blöte HWJ (2005) Monte Carlo study of the site percolation model in two and three dimensions. *Phys Rev E* 72:016126
11. Erdős P, Rényi A (1959) On random graphs. *Publ Math* 6: 290–297
12. Fisher M, Essam JW (1961) Some cluster size and percolation problems. *J Math Phys* 2:609–619
13. Flory P (1941) Molecular Size Distribution in Three Dimensional Polymers I Gelation. *J Am Chem Soc* 63:3083–3090
14. Frisch HL, Sonnenblick E, Vysotsky V, Hammersley JM (1961) Critical percolation probabilities (site problem) *Phys Rev* 124:1023–1022
15. Gamsa A, Cardy J (2007) SLE in the three-state Potts model – a numerical study. *J Stat Mech* P08020
16. Gould H, Tobochnik J (1996) *An Introduction to Computer Simulation Methods*, 2nd edn. Addison-Wesley, Reading, p 444
17. Gould H, Tobochnik J, Christian W (2006) *An Introduction to Computer Simulation Methods*, 3rd edn. Addison-Wesley, Reading
18. Grassberger P (1983) Critical behavior of the general epidemic process and dynamical percolation. *Math Biosci* 63:157–172
19. Grassberger P (1986) On the hull of two-dimensional percolation clusters. *J Phys A* 19:2675–2677
20. Grassberger P (1992) Spreading and backbone dimensions of 2D percolation. *J Phys A* 25:5475–5484
21. Grassberger P (1999) Conductivity exponent and backbone dimension in  $2 - d$  percolation. *Physica A* 262:251–263
22. Grassberger P, Zhang YC (1996) Self-organized formulation of standard percolation phenomena. *Physica A* 224:169–179



23. Grossman T, Aharony A (1984) Structure and perimeters of percolation clusters. *J Phys A* 19:L745–L751
24. Gruzberg I (2006) Stochastic geometry of critical curves, Schramm-Loewner evolutions, and conformal field theory. *J Phys A* 39:12601–12656
25. Havlin S, Nossal R (1984) Topological properties of percolation clusters. *J Phys A* 17:L427–L432
26. Herrmann HJ, Derrida B, Vannimenus J (1984) Superconductivity exponents. In: Herrmann HJ, Derrida B, Vannimenus J (eds) Two- and three-dimensional percolation. *Phys Rev B* 30:4080–4082
27. Herrmann HJ, Hong DC, Stanley HE (1984) Backbone and elastic backbone of percolation clusters obtained by the new method of burning. *J Phys A* 17:L261–L266
28. Hong DC, Havlin S, Herrmann HJ, Stanley HE (1984) Breakdown of Alexander-Orbach conjecture for percolation: Exact enumeration of random walks on percolation backbones. *Phys Rev B* 30:4083–4086
29. Hoshen J, Kopelman R (1976) Percolation and cluster distribution I Cluster multiple labeling technique and critical concentration algorithm. *Phys Rev B* 14:3438–3445
30. Hovi J-P, Aharony A (1996) Scaling and universality in the spanning probability for percolation. *Phys Rev E* 53:235–253
31. Hu CK (1992) Histogram Monte Carlo renormalization-group method for percolation problems. *Phys Rev B* 46:6592–6595
32. Hu CK, Chen J-A, Izmailian S Sh, Kleban P (2000) Recent developments in the Monte Carlo approach to percolation problems. *Comp Phys Comm* 126:77–81
33. Jensen I, Ziff RM (2006) Universal amplitude ratio  $\Gamma^-/\Gamma^+$  for two-dimensional percolation. *Phys Rev E* 74:020101R
34. Kleban P, Ziff RM (1998) Exact results at the two-dimensional percolation point. *Phys Rev B* 57:R8075–R8078
35. Kleban P, Simmons JJH, Ziff RM (2006) Anchored critical percolation clusters and 2D electrostatics. *Phys Rev Lett* 97:115702
36. Kunz H, Souillard B (1978) Essential singularity in the percolation model. *Phys Rev Lett* 40:133–135
37. Leath P (1976) Cluster size and boundary distribution near percolation threshold. *Phys Rev B* 14:5046–5055
38. Lobb CJ, Frank DJ (1984) Percolation conduction and the Alexander-Orbach conjecture in two dimensions. *Phys Rev B* 30:4090–4092
39. Lorenz CD, Ziff RM (1998) Precise determination of the bond percolation thresholds and finite-size scaling corrections for the sc, fcc, and bcc lattices. *Phys Rev E* 57:230–236
40. Lorenz CD, May R, Ziff RM (2000) Similarity of percolation thresholds on the hcp and fcc lattices. *J Stat Phys* 98:961–970
41. Martin-Herrero J (2004) Hybrid cluster identification. *J Phys A* 37:9377–9386
42. Moore C, Newman MEJ (2000) Exact solution of site and bond percolation on small-world networks. *Phys Rev E* 62:7059–7064
43. Moukarzel C (1998) A Fast Algorithm for Backbones. *Int J Mod Phys C* 9:887–895
44. Newman MEJ, Ziff RM (2000) Efficient Monte Carlo algorithm and high-precision results for percolation. *Phys Rev Lett* 85:4104–4107
45. Newman MEJ, Ziff RM (2001) A fast Monte Carlo algorithm for site or bond percolation. *Phys Rev E* 64:016706
46. Osterkamp D, Stauffer D, Aharony A (2003) Anomalous diffusion at percolation threshold in high dimensions of  $10^{18}$  sites. *Int J Mod Phys C* 14:917–924
47. Parviainen R (2007) Estimates of the bond percolation thresholds on the Archimedean lattices. *J Phys A* 40:9253–9258
48. Paul G, Ziff RM, Stanley HE (2001) Percolation threshold, Fisher exponent, and shortest path exponent for four and five dimensions. *Phys Rev E* 64:026115
49. Perlsman E, Havlin S (2002) Method to estimate critical exponents using numerical studies. *Europhys Lett* 58:176–181
50. Pike R, Stanley HE (1981) Order propagation near the percolation threshold. *J Phys A* 14:L169
51. Pinson HT (1994) Critical percolation on the torus. *J Stat Phys* 75:1167–1177
52. Privman V, Hohenberg PC, Aharony A (1991) Universal Critical-Point Amplitude Relations. In: Domb C, Lebowitz JL (eds) Phase transition and critical phenomena, vol 14. Academic Press, New York
53. Quintanilla J, Torquato S, Ziff RM (2000) Efficient measurement of the percolation threshold for fully penetrable discs. *J Phys A* 33:L399–L407
54. Quintanilla J, Ziff RM (2007) Near symmetry of percolation thresholds of fully penetrable disks with two different radii. *Phys Rev E* 76:051115
55. Rammal R, Angles d'Auriac JC, Benoit A (1984) Universality of the spectral dimension of percolation clusters. *Phys Rev B* 30:4087–4089
56. Reynolds P, Stanley HE, Klein W (1980) Large-cell Monte Carlo renormalization group for percolation. *Phys Rev B* 21:1223–1245
57. Rosso M, Gouyet JF, Sapoval B (1985) Determination of percolation probability from the use of a concentration gradient. *Phys Rev B* 32:6063–6054
58. Rozenfeld HD, ben-Avraham D (2007) Percolation in hierarchical scale-free nets. *Phys Rev E* 75:061102
59. Saleur H, Duplantier B (1987) Exact determination of the percolation hull exponent in two dimensions. *Phys Rev Lett* 58:2325–2328
60. Sapoval B, Rosso M, Gouyet J-F (1985) The fractal nature of a diffusion front and relation to percolation. *J Phys Lett Paris* 46:L149
61. Schramm O (1999) Scaling limits of loop-erased random walks and uniform spanning trees. *Israel J Math* 118:221–288
62. Scullard CR, Ziff RM (2006) Predictions of bond percolation thresholds for the kagomé and Archimedean ( $3, 12^2$ ) lattices. *Phys Rev E* 73:045102(R)
63. Scullard CR, Ziff RM (2008) Critical surfaces of general bond percolation problems. *Phys Rev Lett* 100:185701
64. Shchur LN (2000) Incipient Spanning Clusters in Square and Cubic Percolation. In: Landau DP, Lewis SP, Schuettler HB (eds) Springer Proceedings in Physics, vol 85. Springer, Berlin
65. Simmons JJH, Kleban P, Ziff RM (2007) Exact factorization of correlation functions in 2-D critical percolation. *Phys Rev E* 76:041106
66. Simmons JJH, Kleban P, Ziff RM (2007) Percolation crossing formulas and conformal field theory. *J Phys A* 40:F771–F784
67. Smirnov S, Werner W (2001) Critical exponents for two-dimensional percolation. *Math Res Lett* 8:729–744
68. Stauffer D, Aharony A (1994) An Introduction to Percolation Theory, revised 2nd edn. Taylor and Francis, London
69. Suding PN, Ziff RM (1999) Site percolation thresholds for Archimedean lattices. *Phys Rev E* 60:295–283
70. Sykes MF, Essam JW (1964) Exact critical percolation probabilities for site and bond problems in two dimensions. *J Math Phys* 5:1117–1127

71. Tarjan T (1972) Depth-first search and linear graph algorithms. *SIAM J Comput* 1:146–160
72. Temperley HNV, Lieb EH (1971) Relations between the ‘percolation’ and ‘colouring’ problem and other graph-theoretical problems associated with regular planar lattices: some exact results for the ‘percolation’ problem. *Proc R Soc London A* 322:251–280
73. Tiggemann D (2001) Simulation of percolation on massively-parallel computers. *Int J Mod Phys C* 12:871
74. Voigt CA, Ziff RM (1997) Epidemic analysis of the second-order transition in the Ziff-Gulari-Barshad surface-reaction model. *Phys Rev E* 56:R6241–R6244
75. Vollmayr H (1993) Cluster hull algorithms for large systems with small memory requirement. *J Stat Phys* 74:919–927
76. Voss RF (1984) The fractal dimension of percolation cluster hulls. *J Phys A* 17:L373–L377
77. Vyssotsky VA, Gordon SB, Frisch HL, Hammersley JM (1961) Critical Percolation Probabilities (Bond Problem). *Phys Rev* 123:1566–1567
78. Weinrib A, Trugman SA (1985) A new kinetic walk and percolation perimeters. *Phys Rev B* 31:2993–2997
79. Wierman JC (1984) A bond percolation critical probability determination based on the star-triangle transformation. *J Phys A* 17:1525–1530
80. Wu FY (1979) Critical point of planar Potts models. *J Phys C* 12:L645–L650
81. Zabolitzky JG (1984) Monte Carlo evidence against the Alexander-Orbach conjecture for percolation conductivity. *Phys Rev B* 30:4077–4079
82. Ziff RM (1986) Test of scaling exponents for percolation cluster perimeters. *Phys Rev Lett* 56:545–548
83. Ziff RM (1992) Spanning probability in 2D percolation. *Phys Rev Lett* 69:2670–2674
84. Ziff RM (1994) Reply to Comment on Spanning probability in 2D percolation. *Phys Rev Lett* 72:1942
85. Ziff RM (1995) Proof of crossing formula for 2D percolation. *J Phys A* 28:6479–6480
86. Ziff RM (1996) Effective boundary extrapolation length to account for finite-size effects in the percolation crossing function. *Phys Rev E* 54:2547–2554
87. Ziff RM (1998) Four-tap shift-register-sequence random-number generators. *Comput Phys* 12:385–392
88. Ziff RM (2004) Enclosed area distribution in percolation. Talk presented at StatPhys22. [arXiv:cond-mat/0510633](https://arxiv.org/abs/cond-mat/0510633)
89. Ziff RM, Cummings PT, Stell G (1984) Generation of percolation cluster perimeters by a random walk. *J Phys A* 17:3009–3017
90. Ziff RM, Sapoval B (1986) The efficient determination of the percolation threshold by a frontier-generating walk in a gradient. *J Phys A* 19:L1169–L1172
91. Ziff RM, Suding PN (1997) Determination of the bond percolation threshold for the kagomé lattice. *J Phys A* 30:5351–5359
92. Ziff RM, Finch SR, Adamchik VS (1997) Universality of finite-size corrections to the number of critical percolation clusters. *Phys Rev Lett* 79:3447–3450
93. Ziff RM, Lorenz CD, Kleban P (1999) Shape-dependent universality in percolation. *Physica A* 266:17–26
94. Ziff RM, Newman MEJ (2002) Convergence of threshold estimates for two-dimensional percolation. *Phys Rev E* 66:016129
95. Ziff RM, Scullard CR (2006) Exact bond percolation thresholds in two dimensions. *J Phys A* 39:15083–15090

## Books and Reviews

Bollobás B, Riordan O (2006) *Percolation*. Cambridge University Press, Cambridge

## Percolation Phase Transition

MUHAMMAD SAHIMI

Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, USA

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[What is Percolation?](#)

[The Percolation Phase Transition](#)

[Percolation Properties](#)

[Universal Scaling Properties of Percolation](#)

[Variants of Percolation Processes and Their Applications](#)

[Future Directions](#)

[Bibliography](#)

## Glossary

**Accessible bonds or sites** The retained bonds or sites in a percolation lattice that are connected to infinity via at least one path.

**Backbone** The set of (retained) bonds or sites in the sample-spanning percolation cluster that are connected to infinity by more than one independent path.

**Bond percolation** In a bond percolation model, a random lattice is formed from an infinite lattice by retaining each bond of the infinite lattice with probability  $p$ , and deleting the rest. When a bond is retained, so also are its two ends sites.

**Cluster** A connected set of bonds or sites that are retained in a percolation lattice with probability  $p$ .

**Correlation length** The length scale below which a disordered percolation system cannot be regarded as homogeneous.

**Critical exponents** At the percolation thresholds, many percolation properties follow universal power laws, and the exponents that characterize such power laws are called *critical exponents*.

**Percolation transition** The connectivity or geometrical transition between a system in which a sample-spanning cluster of retained sites (or bonds) exists, and one in which no such cluster exists.

**Site percolation** In a site percolation problem, a random lattice is formed from an infinite lattice by retaining each site of the infinite lattice with probability  $p$  and deleting the rest. A bond connecting two retained neighboring sites is also retained.

### Definition of the Subject

This chapter describes the basic percolation problem. We begin by noting that most systems of scientific and practical applications are, at least at some scale, disordered. In such disordered media, then, the connectivity of the elementary or microscopic elements has a profound effect on the media's macroscopic properties. The percolation transition occurs at the percolation threshold, which is the point at which the microscopic elements become connected for the first time, and form a sample-spanning path across the system. Percolation theory aims to describe the effect of the connectivity of the microscopic elements on the effective macroscopic properties of disordered media, particularly in the vicinity of the percolation threshold.

### Introduction

It is well known that *Nature is disordered*. Pure, perfectly characterized, and geometrically immaculate systems are nowhere to be found, except perhaps in books and papers on theoretical physics. Although the concept of an infinite, perfectly periodic crystal lattice is incredibly elegant, it is as remote from experimental reality as possible. Even the best experimentalist who focuses on the purest of substances, exemplified by carefully grown crystals, can hardly ever escape the effects of defects, trace impurities, and finite boundaries. Thus, we must come to terms with *disordered morphology*: variations in the shape and constitution that are often so ill-characterized that we must deem them to be random if we are to describe them, or have any hope of doing so. The morphology of a medium has two major aspects: the *topology*, the interconnectiveness of the system's individual microscopic elements, and the *geometry* – the shapes and sizes of the individual elements.

At the same time, we believe, at least above the quantum mechanical level, in the doctrine of determinism, yet important *continua* exist in which deterministic descriptions of many phenomena are beyond hope. A well-known example is diffusion where, at least over certain length scales, one observes an apparent random process – or *disordered dynamics*. The two types of disorder – morphological and dynamical – are often coupled and present simultaneously. An important example is fluid flow through a porous medium, where the interplay between the disordered morphology of the pore space and the dynam-

ics of fluid motion gives rise to a rich variety of phenomena [6,8].

Research on understanding the macroscopic properties of materials did make remarkable progress by using statistical mechanics and taking advantage of periodic structures, or through the application of such equations as the Boltzmann's equation. However, due to the rather obvious randomness in Nature, and because, in the final analysis, one always must confront the real world which is disordered, it became apparent in the 1960s that a statistical physics of disordered media must be developed to provide methods for deriving macroscopic properties of such media from the laws governing the microscopic world or, alternatively, for deducing their microscopic properties from the macroscopic observations and experimental measurements. Such a statistical physics of disordered media must take into account the effect of *both* the topology and geometry of the media. Although the role of the geometry was appreciated as early as the beginning of the twentieth century, the effect of the topology was ignored for many decades, or was treated in an ad hoc manner, simply because it was thought to be too difficult to be taken into account rigorously.

As the history of science indicates, progress in any research field is not usually made with a constant rate, but rather in a sporadic manner. There are periods when a problem looks so difficult that we do not even know where or how to start attacking it, and periods when some seminal discoveries remove a great obstacle to progress and, thus, enable us to make a great leap forward. An excellent example is the discovery of a new class of superconducting materials by Bednorz and Müller [1], who showed that it is possible to have superconductivity in certain Cu alloys at temperatures  $T > 30$  K. Since their discovery (which brought them the physics Nobel Prize in 1989), the field of superconductivity has advanced remarkably (after not making much progress for decades), so much so that we now have materials that can be superconducting at temperature well above 100 K.

Over the past three decades, the statistical physics of disordered media has been in a rapidly progressing phase, the reason for which is fourfold:

- (i) Rigorous theoretical methods for calculating the macroscopic (average) properties of disordered media have been developed.
- (ii) A large amount of accurate experimental data have been accumulated, thanks to many novel experimental techniques and instruments.
- (iii) Advances in computer technology and computational strategies have enabled us to use precise nu-

merical simulations for obtaining accurate estimates of many properties of disordered materials.

- (iv) The fifth, and perhaps the most important, reason for the rapid development of the statistical physics of disordered media is that, the effect of the interconnectivity of the microscopic elements of disordered media on their macroscopic properties has been understood and appreciated. This has become possible through the development and application of percolation theory, the subject of this section of the Encyclopedia.

### What is Percolation?

Consider electrical (or thermal) conduction through a composite material which is a mixture of conducting and insulating constituents or phases. Assume also that the two phases are randomly distributed. As an idealization, we represent the composite material by a simple-cubic network in which each bond is either conducting with a finite conductivity, or insulating with zero conductivity. Suppose also that we impose a voltage difference between two opposite faces of the network. The question that we ask is: what fraction of the bonds must have a finite conductivity in order for the electrical current to flow through the material, so that it has a nonzero macroscopic conductivity? This is clearly an important question, because its answer tells us, for example, that what (volume) fraction of a composite material, such as carbon black composites that are routinely used in many applications, must be conducting in order for the composite as a whole to be conducting.

Consider a second example. Imagine that the bonds of a simple-cubic network represent the pore throats of a porous medium, e.g., an oil reservoir in porous rock. The pore throats are the narrow passages that connect the pore bodies. Most of the porosity of a porous medium (the void volume fraction in the porous medium) resides in the pore bodies. For brevity, we refer to the pore throats as pores. In reality, no porous medium looks as ordered as a simple-cubic network, but as an idealization the model is useful. Now, suppose that the pores (bonds) are filled with oil, and that there are two wells in the system, one at A on one face, and a second one at B on the opposite face of the network. We try to push the oil out of the network (porous medium) by injecting water into the system at A – the injection well – to produce oil at B – the production well. Oil and water do not mix with each other and, therefore, we assume that each pore is filled with either oil or water. We also assume that water wets the surface of the pores (the wetting fluid), whereas oil does not (the nonwetting

fluid). In many oil reservoirs, such as carbonate oil reservoirs of the Middle East, the opposite is true, but this does not make any difference to our discussion.

When the water is injected and pushed into the reservoir, it tries, due to being the wetting fluid, to find the *smallest pores* that it can reach and expel the oil from it. In reality, the process is more complex than what we are describing, but we ignore all the complications. The displaced oil is produced at well B. The question that we ask is: what fraction of the pores are filled with water when it reaches the production well at B *for the first time* (this is called the *breakthrough point*)? In other words, we would like to know what fraction of the pores lose their oil and, thus, how much oil is produced at well B at the breakthrough point. This is clearly an important question, given that the price of oil is now around \$75/barrel.

### The Percolation Phase Transition

In the example of composite materials described above, if too many bonds (or too much of the materials) are insulating, no macroscopic current will flow through the materials, whereas for sufficiently large number of conducting bonds electrical current does flow in the materials, so that their macroscopic effective conductivity is nonzero. Assume that the fraction of the conducting bonds is  $p$ . Therefore, there must be a minimum or critical value  $p_{cb}$  of  $p$ , such that for  $p \leq p_{cb}$  no electrical current would flow through the material and, therefore, the materials as a whole are insulating, whereas for  $p > p_{cb}$  the materials become conducting.

In the example of displacement of oil by water,  $p$  represents the fraction of pores from which oil has been expelled and replaced by water. Therefore, for  $p \leq p_{cb}$ , water flows only locally, and has not reached the production wall, whereas for  $p > p_{cb}$  water flows between the injection and production wells. Thus, at any given time,  $p$  represents the fraction of the total oil in the reservoir that has been recovered, while  $p_{cb}$  represents its value at the breakthrough point.

Therefore, it should be clear that, in both examples,  $p_c$  signifies a phase transition: for  $p \leq p_{cb}$  there is no sample-spanning path of conducting bonds, or pores filled by water, so that the system is macroscopically disconnected or *closed* (to electrical current or flow of water). But, for  $p > p_{cb}$  the system becomes macroscopically connected. Hence,  $p_{cb}$  is the point at which a *geometrical* phase transition from a disconnected to a connected system takes place. Percolation theory, then, quantifies the effect of the interconnectivity of the microscopic elements of a disordered medium (the conducting elements, or the pores

filled with water) on its macroscopic properties.  $p_{cb}$  is called the *bond percolation threshold* of the network.

We may also formulate the percolation problem in another way. Recall that most of the porosity of a porous medium resides in its pore bodies which connect the pore throats, as several pore throats meet at a pore body. Thus, in the pore network model of a porous medium described above, the nodes or sites of the network, which connect the pore throats or bonds, are the equivalent of the pore bodies. Then, in the example of the displacement of oil by water in a porous medium, the injected water pushes the oil from the sites (into the bonds) toward the production well. Since most of the porosity (the void space available for the fluids) resides in the pore bodies, to obtain a more accurate estimate of the volume of the oil recovered, we ask the question: At the breakthrough point, what fraction of the network's sites (pore bodies) are filled with water? Denoting this fraction by  $p_{cs}$ , it should be clear that it is the analogue of  $p_{cb}$ .  $p_{cs}$  is called the *site percolation threshold* of the network. For all two- and three-dimensional (3D) lattices,  $p_{cs} > p_{cb}$ .

Determination of exact percolation thresholds of many 2D and all the 3D lattices remains an unsolved problem. The article by Wierman describes and discusses the existing exact results for the percolation thresholds. Ziff's article describes highly efficient numerical methods for estimating the percolation threshold and many other properties of percolation lattices.

## Percolation Properties

Some of the most important properties of percolation systems that describe their morphology are as follows. For simplicity, we use  $p_c$  to denote  $p_{cs}$  or  $p_{cb}$ .

1. The *percolation probability*  $P_\infty(p)$  is the probability that, when the fraction of occupied bonds is  $p$ , a given site belongs to the sample-spanning (infinite) cluster of occupied bonds.
2. The *accessible fraction*  $A(p)$  is that fraction of occupied bonds (or sites) that belong to the infinite cluster.
3. The *backbone fraction*  $B(p)$  is the fraction of occupied bonds in the infinite cluster which actually participate in a transport process, such as conduction, since some of the bonds in the infinite cluster are dead-end and do not carry any current. Therefore,  $A(p) \geq B(p)$ .
4. The *correlation length*  $\xi(p)$  is the typical radius of percolation clusters for  $p < p_c$ , and the typical radius of the "holes" above  $p_c$  that are generated by the vacant bonds or sites. For  $p > p_c$ ,  $\xi$  is the length scale over which the system is macroscopically homogeneous.

5. The *average number of clusters of size  $s$*  (per lattice site)  $n_s(p)$  is an important quantity in many of the problems of interest here because it corresponds to, for example, the number of conducting or insulating islands of a given size in a conductor-insulator composite solid.
6. The probability that two sites, one at the origin and another one at a distance  $\mathbf{r}$ , are both occupied and belong to the same cluster of occupied sites, is  $p^2 P_2(\mathbf{r})$ , where  $P_2(\mathbf{r})$  is called the *pair-connectedness function*.
7. The *mean cluster size*  $S$  (also called the *site-averaged cluster number*) is the average number of sites in the cluster that contains a randomly-selected site, and is given by,

$$S = \frac{\sum_s s^2 n_s(p)}{\sum_s s n_s(p)}. \quad (1)$$

Essam [4] showed that  $S$  and the pair-connectedness function  $P_2(\mathbf{r})$  are related through a simple relation:

$$S = 1 + p \sum_{\mathbf{r}} P_2(\mathbf{r}). \quad (2)$$

8. Because a major application of percolation theory has been modeling of transport in disordered materials, and in particular composite solids, we must also consider the effective transport properties of percolation systems, namely, their conductivity, diffusivity, elastic moduli, and dielectric constant. We first consider the conductivity of a two-phase composite material modeled as a two-component network in which each (randomly-selected) bond has a conductance  $g_1$  with probability  $p$  or  $g_2$  with probability  $q = 1 - p$ . It is straightforward to show that the effective electrical (or thermal) conductivity  $\sigma_{\text{eff}}$  of the network is a *homogeneous function* and takes on the following form,

$$\sigma_{\text{eff}}(p, g_1, g_2) = g_1 F(p, h), \quad (3)$$

where  $h = g_2/g_1$ . Due to the assumption of randomness of the material's morphology,  $\sigma_{\text{eff}}$  is invariant under the interchange of  $g_1$  and  $g_2$  (phase-inversion symmetry) and we must, therefore, have

$$\sigma_{\text{eff}}(p, g_1, g_2) = \sigma_{\text{eff}}(q, g_2, g_1), \quad \text{and} \quad (4)$$

$$F(p, h) = h F(q, 1/h).$$

The limit in which  $g_2 = 0$  and  $g_1$  is finite corresponds to a conductor-insulator mixture, already described above. In this case, as  $p \rightarrow p_c$ , more and more bonds are insulating, the conduction paths become very tortuous and, therefore,  $\sigma_{\text{eff}}$  decreases; at  $p_c$  one



has  $\sigma_{\text{eff}}(p_c) = 0$ , since no sample-spanning conduction path exists any more. More generally, the conductance  $g_1$  may follow a certain statistical distribution, which is in fact the case in most systems of practical importance, such as porous materials and composite solids.

The limit in which  $g_1 = \infty$  and  $g_2$  is finite represents a *conductor-superconductor* mixture. All quantum-mechanical aspects of real superconductors are ignored in this definition, and we are concerned only with the effect of the local connectivity of the material on this conductivity. It is clear that the effective conductivity  $\sigma_{\text{eff}}$  of this system is dominated by the superconducting bonds. If  $p < p_c$ , then a sample-spanning cluster of the superconducting bonds does not exist, and  $\sigma_{\text{eff}}$  is finite. As  $p \rightarrow p_c^-$ ,  $\sigma_{\text{eff}}$  increases until a sample-spanning cluster of the superconducting bonds is formed for the first time at  $p = p_c$ , where  $\sigma_{\text{eff}}$  diverges. Note that both limits ( $g_1$  finite and  $g_2 = 0$ , and  $g_1 = \infty$  and  $g_2$  finite) correspond to  $h = 0$ . Therefore, the point  $h = 0$  at  $p = p_c$  is particularly important. The article by Hughes elaborates on these aspects, and provides a full account of the state-of-the-art of this problem.

9. In a similar manner, the elastic moduli of a two-phase composite solid, modeled by a percolation network, are defined. Consider a two-component network in which each bond is an elastic element (a spring or beam) which has an elastic constant  $e_1$  with probability  $p$  or  $e_2$  with probability  $q = 1 - p$ . The limit in which  $e_2 = 0$  and  $e_1$  is finite corresponds to composites of rigid materials and holes (for example, porous solids). In such networks, as  $p \rightarrow p_c$ , more bonds have no rigidity, the paths for transmission of stress or elastic forces become very tortuous and, therefore, the effective elastic moduli  $E$  (Young's, bulk, or shear moduli) decrease; at  $p_c$  one has  $E(p_c) = 0$ . In general, the elastic constant  $e_1$  can be selected from a statistical distribution.

The limit in which  $e_1 = \infty$  and  $e_2$  is finite represents mixtures of rigid-superrigid materials. In this case the effective elastic moduli  $E$  of the system are dominated by the superrigid bonds. If  $p < p_c$ , then a sample-spanning cluster of the superrigid bonds cannot form, and  $E$  is finite. As  $p \rightarrow p_c^-$ , the effective elastic moduli increase until the percolation threshold  $p_c$  of the rigid phase is reached at which a sample-spanning cluster of the superrigid bonds is formed for the first time, and the effective elastic moduli diverge. The article by Duxbury provides a comprehensive discussion of this subject.

10. The effective dielectric constant  $\varepsilon$  of a two-phase insulating composite material, modeled by a percolation network, may also be defined and, in fact,  $\varepsilon$  is closely related to the conductor-superconductor model described above (see, for example, [9]).
11. Finally, the effective diffusivity  $D$  of a porous material can also be defined in a similar manner; see the article [► Conduction and Diffusion in Percolating Systems](#) by Hughes.

### Universal Scaling Properties of Percolation

One of the most important characteristics of percolation systems is their *universal* properties. The behavior of many percolation quantities near  $p_c$  is insensitive to the microstructure (for example, the coordination number) of the network, and to whether the percolation process is a site or a bond problem. The quantitative statement of this universality is that many percolation properties follow power laws near  $p_c$ , and the *critical exponents* that characterize such power laws are universal and depend only on the Euclidean dimensionality  $d$  of the system. We first describe the universal properties of the quantities that characterize the morphology of percolation systems, and then present and discuss those of transport properties.

In general, the following power laws hold near  $p_c$ ,

$$P_\infty(p) \sim (p - p_c)^\beta, \quad (5)$$

$$A(p) \sim (p - p_c)^\beta, \quad (6)$$

$$B(p) \sim (p - p_c)^{\beta_b}, \quad (7)$$

$$\xi(p) \sim |p - p_c|^{-\nu}, \quad (8)$$

$$S(p) \sim |p - p_c|^{-\gamma}, \quad (9)$$

$$P_2(\mathbf{r}) \sim \begin{cases} r^{2-d-\eta}, & p = p_c, \\ \exp(-r/\xi), & \text{otherwise}, \end{cases} \quad (10)$$

where  $r = |\mathbf{r}|$ . For large clusters near  $p_c$ , the cluster size distribution  $n_s(p)$  is described by the following scaling law,

$$n_s \sim s^{-\tau} f[(p - p_c)s^\sigma], \quad (11)$$

where  $\tau$  and  $\sigma$  are two more universal critical exponents, and  $f(x)$  is a scaling function such that  $f(0)$  is not singular. The article by Stauffer elaborates further on these.

Similar power laws are also followed by the transport properties of percolation composites. In particular,

$$\sigma_{\text{eff}}(p) \sim (p - p_c)^t, \quad (12)$$

conductor-insulator composites

$$\sigma_{\text{eff}}(p) \sim (p_c - p)^{-s}, \quad (13)$$

conductor-superconductor composites

$$E(p) \sim (p - p_c)^T, \quad (14)$$

rigid-soft composites

$$E(p) \sim (p_c - p)^{-s}, \quad (15)$$

rigid-superrigid composites.

For length scales  $L < \xi$ , the resistance  $R$  between two end points of a box of linear size  $L$  scales with  $L$  as  $R \sim L^{\tilde{\zeta}}$ . It is not difficult to show that

$$t = (d - 2)\nu + \zeta, \quad (16)$$

where,  $\zeta = \tilde{\zeta}\nu$ . It has been shown [12] that in 2D,  $t = s$ .

The power law that characterizes the behavior of the effective diffusivity  $D(p)$  near  $p_c$  is derived from that of  $\sigma_{\text{eff}}(p)$ , and is shown to be given by

$$D(p) \sim (p - p_c)^{t-\beta}. \quad (17)$$

The implied prefactors in all the above power laws depend on the type of lattice and are *not* universal.

Equations (12) and (13) can be unified by using the two-component resistor network described above. In the critical region, i. e., the region near  $p_c$ , where both  $|p - p_c|$  and  $h = g_2/g_1$  are small, the effective conductivity  $\sigma_{\text{eff}}$  follows the following scaling law [3,11]

$$\sigma_{\text{eff}} \sim g_1 |p - p_c|^t \Phi_{\pm} (h |p - p_c|^{-t-s}). \quad (18)$$

where  $\Phi_+$  and  $\Phi_-$  are two homogeneous functions corresponding, respectively, to the regions above and below  $p_c$ , and are, similar to  $t$  and  $s$ , universal. For any fixed and non-zero  $h$ ,  $\sigma_{\text{eff}}$  has a smooth dependence on  $p - p_c$ . This becomes clearer if we rewrite Eq. (18) as

$$\sigma_{\text{eff}} \sim g_1 h^{t/(t+s)} \Psi \left[ |p - p_c| h^{-1/(t+s)} \right], \quad (19)$$

where  $\Psi(x) = x^t \Phi_+(x^{-t-s}) = (-x)^t \Phi_-((-x)^{-t-s})$ . Since the function  $\Psi(x)$  is universal, the implication of Eq. (19) is that, if one plots  $\sigma_{\text{eff}}/[g_1 h^{t/(t+s)}]$  versus  $|p - p_c| h^{-1/(t+s)}$  for all networks (or randomly-disordered materials) that have the same Euclidean dimensionality, all the results (or measurements) should collapse onto a single universal curve. This provides a powerful tool for estimating the conductivity of a composite for any value of  $h$ , given the conductivities for two other values of  $h$  (by which the universal curve is constructed). Somewhat similar, but more complex, scaling equations can be developed for the elastic moduli, dielectric constant and other properties of percolation composites.

No exact relation is known between the transport and morphological exponents. This is, perhaps, because the transport exponents describe *dynamical* properties of disordered materials and media, whereas the morphological exponents characterize their *static* properties. In general, there is no reason to expect a direct relation between the two.

If two physical phenomena in heterogeneous media that contain percolation-type disorder are described by two different sets of critical exponents, then the physical laws governing the two phenomena must be fundamentally different. Thus, critical exponents help one to distinguish between different classes of problems and the physical laws that govern them. Moreover, since the numerical values of the percolation properties are not universal and vary from one system to another, but the scaling and power laws that they follow near  $p_c$  are universal and do not depend on the details of the system, estimates of the critical exponents for a certain phenomenon are used for establishing the relevance of a particular percolation model to that phenomenon in disordered materials.

### Variants of Percolation Processes and Their Applications

In this section of the Encyclopedia, the theoretical aspects of percolation theory are described first and, then, some well-established applications are described.

It should be clear that a percolation network is created when sites or bonds are blocked or removed and, therefore, the macroscopic connectivity of the system is gradually lost. In the example of the composite materials, the bonds or sites are blocked to the conducting phase. In the example of displacement of oil by water in a porous medium, the bonds or sites are blocked to oil (since it is expelled from such sites or bonds). Therefore, percolation networks are also useful as simple models of any disordered medium in which the connectivity of the medium's microscopic elements influences its macroscopic properties. Moreover, as the articles by Stauffer and Ziff make clear, the main concepts of percolation theory are simple and, therefore, writing computer program for simulating a percolation process is conceptually straightforward and simple, if we do not wish to simulate very large networks. Thus, percolation networks may also serve as a simple tool for introducing students to computer simulation of disordered media. Stauffer and Aharony [10], Bunde and Havlin [2], and Hughes [5] emphasized the theoretical foundations of percolation theory, while Sahimi [7,8,9] described its important applications.

Although percolation in regular lattices – those in which the coordination number  $Z$  (the number of bonds connected to the same site) is the same everywhere – has been extensively invoked for studying the morphology and transport properties of many disordered materials, percolation in continua and in topologically-random networks – those in which the coordination number varies from site to site – are also of great interest, since in many practical situations one may have to deal with such irregular and continuous systems. For example, continuum percolation is directly applicable to characterization and modeling of the morphology and effective transport properties of microemulsions, polymer blends, sintered materials, sol-gel transitions, and many more. The article ► [Continuum Percolation](#) by Balberg describes the advances that have been made in understanding of the percolation effects in continuous systems, and in random networks.

In the percolation phenomena described so far, no correlations between various segments of the system (for example, bonds and/or sites, or their transport properties) were assumed. However, disorder in many important heterogeneous materials is not completely random. There usually are correlations of some extent that may be finite but large. For example, in packing of solid particles, there are short-range correlations. The universal scaling properties of percolation systems with finite-range correlations are the same as those of random percolation, if the length scale of interest is larger than the correlation length. Moreover, if the correlation function  $C(r)$  decays as  $r^{-d}$  or faster, where  $d$  is the Euclidean dimensionality of the system, then the scaling properties of the system are identical with those of random percolation. This is not totally unexpected because even in random percolation, as  $p$  decreases toward  $p_c$ , correlations begin to build up and, therefore, the introduction of any type of correlations with a range shorter than the percolation correlation length  $\xi$  cannot change its scaling properties. In many other cases, e.g., in some disordered elastic materials, there are long-range correlations. The article by Coniglio describes the major differences between percolation in random and correlated systems.

A particular type of percolation model with extended correlations is known as the *bootstrap percolation*. In this problem sites of a lattice are initially randomly occupied. Then, those sites that do not have at least  $Z_c$  nearest-neighbor occupied sites are removed (note that  $Z_c = 0$  is the usual random percolation). The interactions between the sites are short-ranged, but the correlations between them may build up as the distance between two occupied sites also increases. The original motivation for developing this model was to explain the behavior of some dis-

ordered materials in which magnetic impurities are randomly distributed in a host of non-magnetic metals. It is believed that in some of such materials an impurity atom cannot sustain a localized magnetic moment unless it is surrounded by a minimum number of magnetic neighbors. Bootstrap percolation has proven to be a complex problem with a rich variety of unusual properties that are a strong function of the parameter  $Z_c$ . For example, an important question is the nature of the percolation transition in this model. It now appears that for sufficiently high values of  $Z_c \leq Z$  (where  $Z$  is the coordination number of the lattice), the percolation transition is *first-order*, i.e., discontinuous, whereas for low values of  $Z_c$  the transition is continuous and second-order. If the phase transition is first-order, then the percolation threshold of the system is, in fact,  $p_c = 1$ , the sample-spanning cluster is compact, and power laws (5)–(15) are no longer valid. The article ► [Bootstrap Percolation](#) by De Gregorio et al. describes this important area of percolation problems.

Over the past three decades percolation theory has been applied to modeling of a wide variety of phenomena in disordered media and systems. It is impossible to describe and discuss all such applications. In this section of the Encyclopedia, several well-established and well-understood applications are described and discussed.

The article ► [Invasion Percolation](#) by Knackstedt and Paterson describes in detail application of the percolation model to two-phase fluid flow in porous media, a simple example of which was already described above. Since, as described above, one fluid is injected into a porous medium – that is, it invades the medium – in order to displace the second fluid, this particular model is usually known as the *invasion percolation*. Other aspects of the application of percolation in problems on fluid flow through porous media are described in the article ► [Percolation in Porous Media](#) by King and Masihi.

It appears that percolation provides a powerful tool for modeling of the effect of the connectivity of fractures and faults on fluid flow and transport properties of rock, a highly complex set of phenomena. Thus, in their article ► [Percolation, and Faults and Fractures in Rock](#), Adler et al. describe the recent application of percolation to this important problem.

The article ► [Networks, Flexibility and Mobility](#) by Thorpe describes recent advances on generalization of the percolation model, and its application to modeling of network glasses and proteins. The question of the rigidity of such materials is addressed. Thorpe's article is related to Duxbury's review ► [Elastic Percolation Networks](#) which describes elastic properties of percolation networks and their applications.

An important and well-established application of percolation is to modeling of the rheology of polymers and gels, particularly in the vicinity of the gelation point. Several variants of the percolation models have been developed in order to address this important problem. The article by Sahimi describes the advances that have been made in this area.

Finally, the article ► [Percolation in Complex Networks](#) by Cohen and Havlin describes application of percolation to problems in complex networks, particularly scale-free networks. They show how the concepts of percolation can be used to study not only the robustness and vulnerability of random networks, but also such problems as immunization and epidemic spreading in populations and computer networks, communication paths, and fragmentation in social networks.

### Future Directions

Theoretically, most aspects of percolation are well-understood. Exact values of most of the critical exponents (in 2D), or their very accurate numerical estimates (in 3D), are known. Exact values of the bond and site percolation thresholds of several 2D lattices are also known, as are very accurate numerical estimates of the percolation thresholds of many 3D lattices, although we still do not know the exact value of, for example, the site percolation threshold of the square lattice. Thus, theoretically, the grand challenge is to develop general methods for obtaining the exact percolation thresholds of 3D lattices, and the exact values of the critical exponents for 3D systems, although the latter challenge may well be beyond reach.

Therefore, aside from the theoretical challenges described above, most of the future work on percolation will be concentrated on its applications to problems of practical importance, examples of which are described in this section of the Encyclopedia.

### Bibliography

1. Bednorz JG, Müller KA (1986) Possible high  $T_c$  superconductivity in the Ba-La-Cu-O system. *Z Phys B* 64:189
2. Bunde A, Havlin S (eds) (1996) *Fractals and Disordered Systems*, 2nd edn. Springer, Berlin
3. Efros AL, Shklovskii BI (1976) Critical behavior of conductivity and dielectric constant near the metal-non-metal transition point. *Phys Status Solidi B* 46:475
4. Essam JW (1972) Percolation and cluster size. In: Domb C, Green MS (eds) *Phase Transitions and Critical Phenomena*, vol 2. Academic Press, London, p 197
5. Hughes BD (1996) *Random Walks and Random Environments*, vol 2. Oxford University Press, London
6. Sahimi M (1993) Flow phenomena in rocks: from continuum models to fractals, percolation, cellular automata, and simulated annealing. *Rev Mod Phys* 65:1393
7. Sahimi M (1994) *Applications of Percolation Theory*. Taylor and Francis, London
8. Sahimi M (1995) *Flow and Transport in Porous Media and Fractured Rock*. VCH, Weinheim
9. Sahimi M (2003) *Heterogeneous Materials I & II*. Springer, New York
10. Stauffer D, Aharony A (1994) *Introduction to Percolation Theory*, 2nd edn. Taylor and Francis, London
11. Straley JP (1976) Critical phenomena in resistor networks. *J Phys C* 9:783
12. Straley JP (1977) Critical exponents for the conductivity of random resistor lattices. *Phys Rev B* 15:5733

## Percolation and Polymer Morphology and Rheology

MUHAMMAD SAHIMI

Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Percolation Models of Polymerization and Gelation](#)

[Branched Polymers](#)

[Rheology of Critical Gels](#)

[Resistor and Elastic Percolation Networks](#)

[Nearly-Critical Chemical Gels: Comparison of the Data with the Percolation Models](#)

[Physical Gels: Comparison of the Data with the Percolation Model](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Branched polymers** Large polymers below the gel point with radii larger than the correlation length

**Chemical gels** Gel networks in which the monomers are co-valently bonded.

**Critical gels** The gel networks at the gel point

**Elastic networks** Networks in which each bond is an elastic element, such as a Hookean spring.

**Gel point** The point at which the critical gel network is formed for the first time.

**Lattice animals** Large percolation clusters below the percolation threshold with radii larger than the correlation length

**Physical gels** Gel networks in which the monomers or particles are connected through weak association.

**Relaxation time spectrum** The distribution of relaxation times that describes the linear viscoelastic behavior of liquids and solids.

**Resistor networks** Networks in which each bond is a resistor with a given conductivity.

**Rigidity percolation** Percolation networks in which each uncut bond is a Hookean spring, and there are no angle-changing forces.

**Sol** The solvent + finite polymer clusters below the gel point

### Definition of the Subject

Elastic percolation networks described in the article by Duxbury (see also below), random resistor networks described in the articles by Hughes and Balberg, and a few continuum models [58] of heterogeneous materials provide a comprehensive understanding of their transport properties. The purpose of this chapter is to describe and discuss applications of such percolation models to predicting the structure and rheology of an important class of disordered materials, namely, polymers and gel networks, and test their validity by comparing their predictions with the relevant experimental data.

The formation of the polymeric materials that we consider in this chapter is characterized by the existence of a percolation-type transition point; see below. The rigidity and linear elastic properties of disordered materials, including polymers, that are far from their percolation threshold are well-described and predicted by mean-field theories, such as the effective-medium approximation [33,41]. However, the effective properties of polymeric materials that are near the percolation threshold deviate greatly from the predictions of mean-field theories and other approximations. It is the description of various properties of such polymeric materials that is best done by the percolation models, and is the focus of this chapter.

### Introduction

Polymeric materials have wide applications in many branches of science and technology. In addition, they have many interesting, and in many cases unusual, properties that justify their study. One is that their relaxation modes are described by a wide spectrum, which provides clues to their structure (see below). Each mode is associated with a particular “event” or motion. In particular, motion of clusters of monomers or molecules is associated with the long modes, with the longest relaxation modes being due to the very large clusters. Such clusters are formed either

by the formation of *chemical bonds* or *chemical crosslinking* between the monomers and also between monomers and small molecules or clusters which, when large enough, lead to *phase-separation*, or by *physical association* at the molecular or particulate level. It is the formation of such large clusters that is the root cause of an important class of phase transitions, namely, the *liquid–solid transition* (LST).

The significance of the LST cannot be over-emphasized, as it occurs in a wide variety of problems of practical importance. At the most basic level, it is important to be able to predict the point at which the LST occurs. The knowledge is sometimes necessary in order to design better polymer processing operations. An example is injection molding of semicrystalline polymers, the quality of the surface of which is a strong function of the location of the LST point. In other applications the knowledge of the LST point is necessary in order to *postpone* it or *avoid* it altogether. Since near the LST point the system is a mixture of a liquid and solid clusters, one may be able to design a wide variety of materials by changing the volume fraction of each phase and, therefore, study of the LST is important. At the same time, crosslinked polymers near the LST point are good adhesives, and have also been used as materials for membranes, absorbers, and many other applications.

In addition to chemists and chemical and polymer engineers who have traditionally studied polymers and their properties, a seminal work in the early 1970s attracted the attention of physicists, and opened the way to the application of modern methods of statistical mechanics to the study of polymers. De Gennes [71] demonstrated that there is a close connection between linear polymers – those in which the monomers have functionality or coordination number  $Z = 2$  – and a statistical mechanical model, namely, the  $n$ -vector model. Clearly, no two monomers occupy the same point in space. If, in addition, there is no closed loop in the structure of the polymer, then, the result is a *linear polymer* which, as de Gennes showed, corresponds to the limit  $n \rightarrow 0$  of  $n$ -vector model. The most suitable model for such polymers is the path of a self-avoiding walk in which a particle performs a random walk in space with the restriction that it never visits any point more than once (so that loop formation is avoided). De Gennes’ discovery made it possible to apply modern methods of statistical mechanics, and in particular the renormalization group theory and the scaling concepts, to the study of linear polymers and, later, to branched polymers and gels.

The original work of de Gennes was restricted to linear polymers. However, if the monomers have functionalities



$Z > 2$ , so that each of them may be connected with more than two neighboring monomers, then at least two other classes of polymers can be obtained:

- (i) If the reaction time  $t$  is relatively short and below, but close to, a characteristic time  $t_g$ , then one obtains branched polymers in the solution, usually called *sol*, that form a viscous solution. It is called a *sol* because it is soluble in good solvents. Such branched polymers are large but finite clusters of monomers.  $t_g$  is called the *gelation time*.
- (ii) If, however, the reaction time is larger than  $t_g$ , an infinitely large solid network of reacted monomers appears which is usually called a *chemical gel*, or simply a gel. There is clearly a LST as one passes from the sol to the gel phase, and the process, called *gelation*, has been described by the percolation models.

The gel can only swell, but not dissolve, in a solvent, even if finite clusters of reacted monomers still exist in the system. The point at which the gel network appears for the first time – which is, in fact, the point that signals the LST – is called the *gel point* (GP).

The gel network has interesting structural, mechanical, and rheological properties which are described in this chapter. Most of us are already familiar with such sol–gel transformations in our daily lives, since we all know about milk-to-cheese transition, pudding, gelatine, etc. However, materials that contain gels, or use their specific properties, are numerous. In addition to the examples mentioned above, another important example is the eye humor. Moreover, gels play an important role in laboratory technology (e.g., gel chromatography), in the fabrication of a wide variety of products, such as glues, cosmetics, contact lenses, etc., and in food technology. In addition, the sol–gel transition is a general phenomenon that has been utilized for producing a variety of ceramic materials [19].

Chemical reactions are responsible for the interconnectivity of the monomers in chemical gels. In general, there are three types of chemical gelation:

- (i) *Polycondensation*, in which polymerization begins with either bifunctional units A–A or trifunctional ones  $B_3$ , or more generally  $Z$ -functional units  $B_Z$ . The A units are linked with the B units, with each elementary reaction being accompanied by the elimination of a molecule between units of A and B. Thus, a polymer network is formed in which the polymer chains are terminated by either A or B. No two units of the same class can participate in a reaction with each other and, therefore, there is always exactly one

bifunctional unit between polyfunctional units in the polymer network.

- (ii) *Vulcanization* begins with long linear polymer chains in a solution. The chains are then crosslinked by small units. An example is rubber, the elasticity of which is due to the introduction of S–S bonds between polyisoprene chains. Only a small number of bonds are needed to crosslink the chains and form an interconnected polymer network. De Gennes [70] argued that for vulcanizing polymers – those with high molecular weights – there exists only a very narrow region near the GP in which the percolation model may be applicable. In other words, such polymers exhibit the Flory–Stockmayer-type behavior [34,63] which is of the mean-field type. For this reason, we ignore vulcanization in this chapter.
- (iii) *Additive polymerization* is similar to polycondensation. The initial solution contains two types of units. The A=A units that are bifunctional when the double bond opens, and the B=D=B units that are quadrifunctional when the two double bonds open independently of each other. If the reaction polymerizes A=A units, one obtains A–A–A–A–... chains, whereas reaction between the A units and the B=D=B units reticulates the network. The length of the chains between two reticulation points is not fixed, but depends crucially on the initiation process and on the relative concentrations of the bi- and quadrifunctional units.

In addition to such chemical gels, one may also have *physical gels* in which the monomers or particles are attached to each other by relatively weak and *reversible* association, or by such physical processes as entanglement. A well-known example is silica aerogel. Another example is a solution of gelatin in water below a certain critical temperature where a coil-to-helix transition takes place, and bonds appear to form by winding of helices of two adjacent chains. Such physical gels can be made and also destroyed by thermal treatment. Other important examples include liquid crystalline polymers at their nematic-to-smectic transition, suspensions and emulsions at the percolation threshold, partially crystalline polymers, and microphase-separating block copolymers.

In this chapter we describe modeling of structural, mechanical and viscoelastic properties of the sol and gel phases, especially near the GP. Modeling of gel network formation was pioneered by Flory [34] and Stockmayer [63], whose theory is essentially equivalent to percolation on the Bethe lattice, an endlessly branching net-

work without any closed loops. Stauffer [62] and de Gennes [72] emphasized the importance of the deviations from the Bethe lattice solution of Flory and Stockmayer, and proposed to replace it by percolation on three-dimensional (3D) lattices. This aspect of the problem, which can be described by a random percolation model or a variant of it, is now well-understood. De Gennes [72] also proposed that the elastic and viscoelastic properties of the gel and sol phases can be described by appropriate random resistor network models (see below). His suggestion was widely accepted for a long time, and was utilized for interpreting the experimental data. It was recognized in the 1980s that, while de Gennes' suggestion may be applicable to certain classes of polymeric materials, more general models are needed for several other important classes of such materials. This realization gave rise to the development of the elastic percolation models that are described in the article by Duxbury. We shall come back to this point shortly.

### Percolation Models of Polymerization and Gelation

To see the connection between the sol–gel transition and the percolation model, consider a solution of molecules or monomers with functionality  $Z \geq 3$ . Suppose, for simplicity, that the monomers occupy the sites of a periodic lattice. With probability  $p$ , two nearest-neighbor monomers (sites) react and form a chemical bond between them. If  $p$  is small, only small polymers (clusters of reacted monomers) are formed. As  $p$  increases, increasingly larger polymers with a broad size distribution are formed. This mixture of clusters of reacted monomers and the isolated unreacted monomers represents the sol phase. For  $p > p_c$ , where  $p_c$  is a characteristic value that depends on the functionality  $Z$  (the number of nearest neighbors of a monomer in the lattice), an infinite cluster of reacted monomers is formed which represents the gel network described above. The gel network at the GP is usually called the *critical gel*. Near the GP the gel usually coexists with the sol such that the finite polymers are trapped in the interior of the gel. For  $p \rightarrow 1$ , almost all the monomers react, and the sol phase disappears completely. Thus,  $p_c$  signals a *connectivity* transition: For  $p > p_c$ , an infinite cluster (the gel network), together (possibly) with a few finite-size clusters, exist and, thus, the system is mainly a rigid gel. The fraction of chemical bonds formed at the GP (which is related to the fraction of the reacted monomers at the GP) is obviously the analogue of the bond percolation threshold. Thus, it should be clear that the formation of branched polymers and gels is very similar to a percolation process.

In the earlier days of modeling the sol–gel transition, it was generally believed that the properties of the polymeric materials at the GP are independent of the structural details of the materials. But, despite decades of research, this is still an unproven hypothesis for the critical gel. In addition, the monomers do not react randomly. There are usually some correlations in the way the monomers react with one another.

### Structural Properties of Branched Polymers and Gels

Studies of the sol–gel transition usually proceed by measuring the time evolution of the rheological (e. g., the viscosity) or mechanical properties (e. g., the elastic moduli) during the chemical reaction that leads to gelation, assuming that the experimental parameter – time or frequency – and the theoretical one – the number of crosslinks – are linearly related in the vicinity of the GP. If true, then, near the GP,  $|p - p_c|$  is a measure of the distance from the GP. All the properties of nearly critical gels (i. e., those that are very near the GP) can be expanded in powers of  $|p - p_c|$ . This is completely similar to, for example, the vapor–liquid phase transition for which all the properties of the system near the critical temperature  $T_c$  can be expanded in powers of  $|T - T_c|$ . But, as the distance from the GP increases, the expansions breakdown.

Experimental rheological measurements are usually performed by using a cone and plate rheometer, or by the more accurate magnetic sphere rheometer. The ranges of shear rates, deformations, and times of measurements of such devices allow the determination of the steady-state zero-shear viscosity and the steady-state linear elastic moduli up to the vicinity of the LST at the GP, but it has proven to be almost impossible to do such measurements at the GP (see below).

The correlation or connectivity length  $\xi$ , which represents the typical size of the branched polymers below  $p_c$ , diverges as  $p_c$  is approached according to the power law

$$\xi \sim |p - p_c|^{-\nu}, \quad (1)$$

which is completely similar to  $\xi_p$ , the correlation length of percolation clusters. Below  $p_c$ , however, the polymers with radii much larger than  $\xi$  have completely different characteristics than those with a typical radius  $\xi$ . Therefore, we describe and discuss such polymers separately, and refer to them as the *branched polymers* to distinguish them from gel networks. Above the GP the correlation length of the polymers is taken as the mesh size of the gel network. For any length scale greater than  $\xi$  the gel network is homogeneous.

### Scaling Properties of the Structure of Nearly Critical Gels

Several important structural properties of branched polymers and gel networks can be measured directly or indirectly. The *gel fraction*  $f_g(p)$  is the fraction of the monomers that are in the gel network, and is measured by simply weighing the gel at different times during the polymerization process. Clearly,  $f_g(p) > 0$  only if  $p > p_c$ . As far as the analogy with the percolation model is concerned,  $f_g$  is the analogue of percolation fraction or percolation probability  $P(p)$  (see the article by Sahimi). Of particular interest to us is the behavior of  $f_g(p)$  near  $p_c$ . In this region,

$$f_g(p) \propto (p - p_c)^\beta. \quad (2)$$

The number distribution of the polymers, i.e., the probability  $Q(s, \epsilon)$  that a polymer in the sol phase contains  $s$  monomers at a distance  $\epsilon = |p - p_c|$  from the GP, is clearly the analogue of the cluster size distribution  $n_s$  for percolation clusters (see the article by Sahimi). Thus, we may write

$$Q(s, \epsilon) \sim s^{-\tau} h_1(\epsilon s^\sigma), \quad (3)$$

where  $h_1$  is a universal scaling function. Instead of writing the distribution in terms of  $s$ , we may write down a power law for the cluster *mass distribution*,  $N(M)$ , in terms of the molecular weight  $M$  of the finite polymers. At the GP, one has,

$$N(M) \propto M^{-\tau}, \quad p = p_c \quad (4)$$

Then, near the GP, one has

$$N(M) \propto M^{-\tau} h_2(M/M_z), \quad (5)$$

which is completely similar to Eq. (3), where  $h_2$  is another scaling function, closely related to  $h_1$ .

Using  $Q(s, \epsilon)$ , we define two distinct mass averages. One, the *weight-average molecular weight*, is defined by

$$M_w = \frac{\int s^2 Q(s, \epsilon) ds}{\int s Q(s, \epsilon) ds} \sim |p - p_c|^{-\gamma}, \quad (6)$$

where  $\gamma + 2\beta = \nu d$ , with  $d$  being the dimensionality of the material. In the analogy with the percolation model,  $M_w$  is the analogue of the mean-cluster size (see the articles by Stauffer and Sahimi). In the polymer literature  $M_w$  is also called the *degree of polymerization*. The second mass average is defined by

$$M_z = \frac{\int s^3 Q(s, \epsilon) ds}{\int s^2 Q(s, \epsilon) ds} \sim \epsilon^{-1/\sigma} \sim |p - p_c|^{-1/\sigma}, \quad (7)$$

where  $M_z$  is the same quantity as in Eq. (5), and,  $\sigma = (\tau - 2)/\beta$ . Note, however, that the average

$$\langle M \rangle = \frac{\int s Q(s, \epsilon) ds}{\int Q(s, \epsilon) ds}$$

does *not* diverge at the GP. Note also that we may also express  $M_w$  in terms of  $\xi$ :

$$M_w \propto \xi^{\gamma/\nu}. \quad (8)$$

Recall (see the article by Stauffer) that percolation theory predicts that for 3D systems,  $\nu \simeq 0.89$ ,  $\beta \simeq 0.41$ ,  $\tau \simeq 2.18$ ,  $\sigma \simeq 0.46$ , and  $\gamma \simeq 1.82$ . The mean-field theory of Flory and Stockmayer predicts the same type of power laws, but with,  $\beta = \gamma = 1$ ,  $\nu = \sigma = 1/2$ , and  $\tau = 5/2$ .

At the GP the gel network is *not* homogeneous, but is a self-similar fractal object with a fractal dimension  $D_f$  that in  $d$ -dimensions is given by

$$D_f = d - \beta/\nu = d(\tau - 1)^{-1}, \quad (9)$$

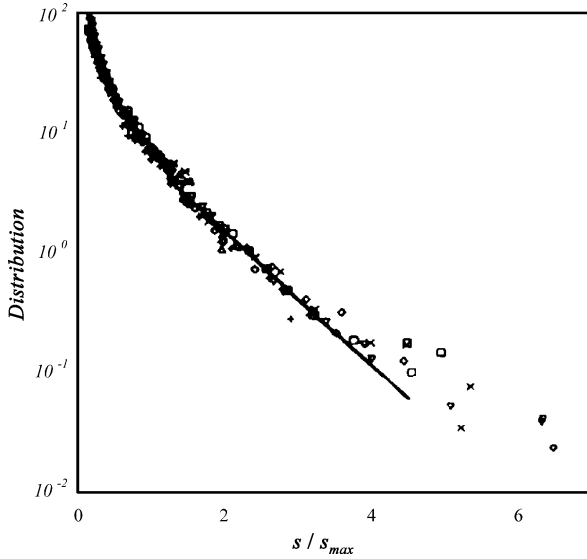
Using numerical estimates of  $\nu$  and  $\beta$  for the 3D percolation model, we obtain,  $D_f \simeq 2.53$ . On the other hand, the Flory–Stockmayer theory predicts that,  $D_f = 4$ , which is unphysical since  $D_f$  cannot be larger than 3.

### Comparison of the Data with the Percolation Model

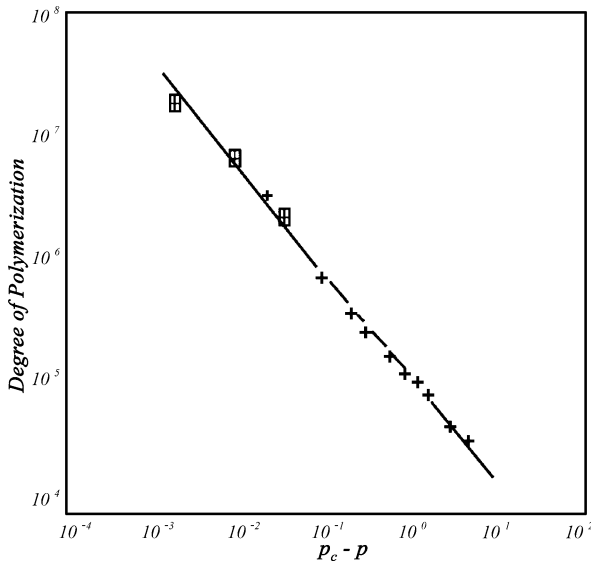
Since a main prediction of percolation is the existence of universal critical exponents and the fractal dimension  $D_f$ , and because the numerical value of any polymer property, such as its average molecular weight or the location of the GP, is not universal and depends on the structure of the polymer, we focus on a comparison between the measured exponents, such as  $\beta$ ,  $\gamma$  and  $\sigma$ , and the predictions of the percolation model.

In their experiments with irradiated polystyrene solution in cyclopentane, Leibler and Schosseler [44] coupled gel permeation chromatography and light scattering to deduce the polymer size distribution which provides a direct means of measuring the exponent  $\tau$ . Figure 1 presents their measurements from which one obtains,  $\tau \simeq 2.3 \pm 0.1$ , close to the percolation prediction of 2.18. Lapp et al. [43] further checked this result by carrying out similar experiments in a system made by chemical end-linking of polydimethylsiloxane, and Patton et al. [51] performed experiments in a system in which polyester was made by bulk condensation polymerization. The measurements of both groups were consistent with the value of  $\tau$  predicted by the percolation model.

Adam et al. [3] tested the validity of the percolation model based on Eq. (6). They carried out static



**Percolation and Polymer Morphology and Rheology, Figure 1**  
Normalized polymer size distribution as a function of polymer size  $s$ . Percolation theory predicts the slope to be  $1 - \tau \simeq -1.3 \pm 0.1$  (after Leibler and Schossler [44])



**Percolation and Polymer Morphology and Rheology, Figure 2**  
Dependence of degree of polymerization  $M_w$  on  $p_c - p$  for a polyurethane sol. The slope of the curve is  $-\gamma$  (after Adam et al. [3])

light scattering measurements on a polyurethane sol. Candau et al. [20] performed their experiments on polystyrene systems crosslinked with divinylbenzene. Figure 2 displays the results of Adam et al. [3] from which one obtains,  $\gamma \simeq 1.71 \pm 0.06$ , only 5% smaller than the percolation

prediction of 1.82. A similar estimate of  $\gamma$  was reported by Candau et al. [20]. On the other hand, one can also express the weight-average molecular weight in terms of the gel fraction near the GP,  $M_w \sim f_g^{-\gamma/\beta}$  and, thus, a plot of  $\log(M_w)$  versus  $\log(f_g)$  yields an estimate of  $\gamma/\beta$ . Schmidt and Burchard [61] carried out anionic copolymerization of divinylbenzene with styrene and obtained both branched polymers (see below) and gels. Light scattering was used to measure the various properties of interest. When Schmidt and Burchard [61] plotted  $\log(M_w)$  versus  $\log(f_g)$ , they obtained a straight line with the slope,  $\gamma/\beta \simeq 4.5$ , in good agreement with the percolation prediction,  $\gamma/\beta \simeq 4.44$ .

### Branched Polymers

After a polymer is formed by crosslinking, the experimentalist usually analyzes its structure by diluting it in a good solvent. As mentioned above, branched polymers in a dilute solution of a good solvent may swell and have a radius *larger* than their extent at the end of the crosslinking. Thus, it is important to consider both typical polymers, which we already described above, and the swollen ones which we now consider.

Thus, consider a swollen branched polymer in a good solvent with a radius larger than the polymer correlation length  $\xi$ . The structural properties of such branched polymers are described by *lattice animals*, which are, in fact very large percolation clusters below the percolation threshold. Their radii are *larger* than the percolation correlation length  $\xi_p$ . But, the interesting and important point is that, although lattice animals are simply very large percolation clusters below  $p_c$ , their statistics are completely *different* from those of percolation clusters.

### Statistics of Lattice Animals

To better understand the difference between percolation clusters and lattice animals, let us first define a few key statistics of lattice animals. Suppose that  $A_s(p)$  is the average number (per lattice site) of the clusters, and  $a_{sm}$  the total number of geometrically different configurations for a cluster of  $s$  sites and perimeter  $m$ . Thus,  $A_s(p) = \sum_m a_{sm} p^s (1-p)^m$ . The asymptotic behavior of  $A_s(p)$  for large values of  $s$  is described by the power law

$$A_s(p) \sim s^{-\theta}, \quad (10)$$

where  $\theta$  is a universal exponent, independent of the coordination number of the lattice. Moreover, for large values of  $s$  a fractal dimension  $D_f$ , defined by

$$s \sim R^{D_f}, \quad (11)$$

describes the structure of the animals or the branched polymers, where  $R$  is the radius of the lattice animal. Note that the fractal dimension  $D_f$  is *distinct* from that of critical gels given by Eq. (9). Lubensky and Isaacson [46] and Family and Coniglio [31] showed that the exponents  $\theta$  and  $D_f$  are not related to any of the percolation exponents defined above. Moreover, Parisi and Sourlas [50] showed that

$$\theta = \frac{d-2}{D_f} + 1, \quad (12)$$

and that

$$D_f = 2, \quad d = 3. \quad (13)$$

One key difference between lattice animals and percolation clusters is that, the exponents  $\theta$  and  $D_f$  are defined for *any*  $p < p_c$  (recall that the percolation exponents are defined for  $p \simeq p_c$ ), so long as  $R$ , the animals' radius, is much larger than the correlation length  $\xi$ . For this reason, they are called *non-critical* exponents.

We may also define a pair correlation function  $C(r)$ , i. e., the probability that two monomers (or sites) that are separated by a distance  $r$  belong to the same polymer (or cluster). For a  $d$ -dimensional branched polymer and large  $r$ , we expect the correlation function to decay as

$$C(r) \sim r^{D_f-d}. \quad (14)$$

The Fourier transform of  $C(r)$  is proportional to the scattered intensity  $I(q)$  in an X-ray or a neutron scattering experiment, where  $q$  is the magnitude of the scattering vector, given by,  $q = (4\pi/\xi) \sin(\theta/2)$ , with  $\theta$  being the wavelength of the radiation scattered by the material through an angle  $\theta$ . Thus, by Fourier transforming Eq. (14), one obtains

$$I(q) \sim q^{-D_f}. \quad (15)$$

In practice, however, polymer solutions are almost always polydisperse and contain polymers of all sizes with radii that may be smaller or larger than the correlation length  $\xi$ . Thus, one must define *average* properties, where the averaging is taken over the polymer size distribution. An average polymer radius is defined by,

$$\langle R \rangle = \frac{\sum_s s^2 R(s) Q(s, \epsilon)}{\sum_s s^2 Q(s, \epsilon)}, \quad (16)$$

which, when combined with Eqs. (3) and (11), yields a relation between  $s$  and  $\langle R \rangle$  [26]:

$$s \sim \langle R \rangle^{D_f(3-\tau)}, \quad (17)$$

so that, in analogy with Eq. (11), an *effective fractal dimension*,  $D_f^e = D_f(3 - \tau)$ , may be defined. Note that, Eq. (17) mixes the branched polymers fractal dimension  $D_f$  with the gel exponent  $\tau$ . If a percolation description of polymerization is correct (which the experiments described above confirmed it to be the case), we should have,

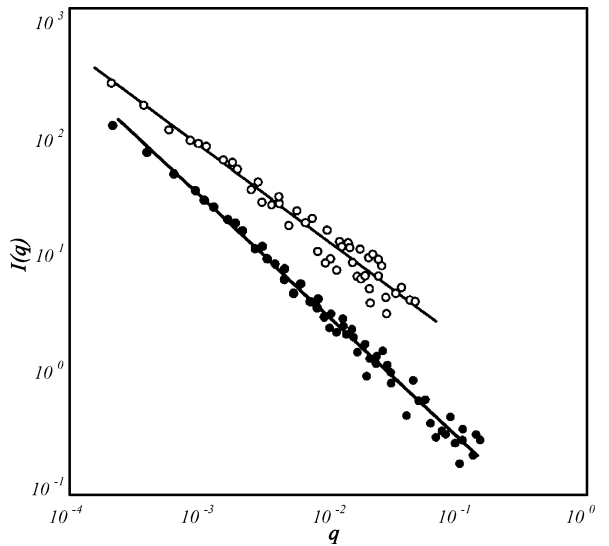
$$D_f^e \simeq 1.64, \quad d = 3, \quad (18)$$

indicating that the effective fractal dimension is smaller than that of a single branched polymer. Because an effective fractal dimension has been defined for a dilute polydisperse polymer solution, the scattering intensity for the same solution should also be modified to

$$I(q) \sim q^{-D_f(3-\tau)}. \quad (19)$$

### Comparison of the Data with the Lattice Animal Model

Experimental evidence for Eq. (13) is actually provided through Eqs. (15) and (19). Bouchaud et al. [18] carried out small-angle neutron scattering experiments on a monodisperse polyurethane sample and measured the scattered intensity as a function of  $q$ . Figure 3 presents their data for the polymer from which one obtains,  $D_f = 1.98 \pm 0.03$ , in excellent agreement with Eq. (19). Bouchaud et al. [18] also synthesized a natural polydisperse polyurethane sample and carried out small-angle neutron scattering on a dilute solution of it. Their data



**Percolation and Polymer Morphology and Rheology, Figure 3**  
Small-angle neutron scattering data for branched polymers. The *upper curve* is for a polydisperse polymer solution with a slope  $-1.6$ . The *lower curve* is for a single polymer in a good dilute solvent with a slope  $-1.98$  (after Bouchaud et al. [18])



yielded the estimate,  $D_f^c \simeq 1.6 \pm 0.05$ , in good agreement with the theoretical prediction given by (18).

Adam et al. [3] carried out static light scattering experiments with dilute polydisperse polyurethane solutions and reported that,  $D_f^c \simeq 1.62 \pm 0.08$ , again in good agreement with (18). Leibler and Schosseler [44] measured the average radius of polystyrene, crosslinked by irradiation by elastic light scattering and found that,  $D_f^c \simeq 1.72 \pm 0.09$ , relatively close to the estimate obtained from Eq. (18). Patton et al. [51] performed both quasi-elastic and elastic light scattering experiments on branched polyesters and reported that,  $D_f^c \simeq 1.52 \pm 0.1$ , somewhat lower than the prediction (18), but still consistent with it.

### Rheology of Critical Gels

One way of understanding what happens to the polymers as the GP is approached is through rheological experiments. Such experiments have been described in details by Winter and Mours [67] and, therefore, we provide only brief description of them and discuss their implications. In practice one imposes a small step shear strain  $\epsilon_{zx}$  on a sample near the GP and measures the shear stress  $\sigma_{zx}(t)$  as a function of time  $t$ . The key property is then the shear stress relaxation function,  $G(t) = \sigma_{zx}/\epsilon_{zx}$ , which is also referred to as the relaxation modulus. Unlike the elastic moduli,  $G(t)$  can be measured for both liquids and solids and, therefore, it is a very useful property for studying the sol-gel transition and, more generally, any LST. Note that  $t$ , the time of crosslinking reaction, corresponds to the extent  $p$  of the reaction, the key property in percolation theory.

What happens to  $G(t)$  as the GP is approached? The stress relaxes quickly at the early stages of crosslinking. As more chemical bonds are formed between the monomers, however,  $G(t)$  stretches out further, since the relaxation process requires longer times. Exactly at the GP (and, more generally, at any LST point), the material is neither a liquid nor a solid yet (it has a tenuous fractal structure). The relaxation modulus follows a power law:

$$G(t) = G_0 t^{-n}, \quad t_0 < t < \infty \quad (20)$$

where  $G_0$  is the gel stiffness. The parameters  $G_0$ ,  $n$ , and  $t_0$  all depend on the material structure at the GP. The exponent  $n$  is closely related to the exponents that characterize the power-law behavior of viscosity of the sol and the elastic moduli of the gel network near the GP (see below).

As the polymerization proceeds further and the GP  $p_c$  is passed, the material becomes a solid which has a finite relaxation modulus at long times, usually referred to as the

equilibrium modulus  $G_e$ :

$$G_e = \lim_{t \rightarrow \infty} G(t). \quad (21)$$

Under such conditions, the stresses can no longer relax completely.

### The Relaxation Time Spectrum

Since the time dependence of a macroscopic relaxation process is always indicative of the underlying microscopic dynamics, one may look for kinetic equations that correctly describe the time-dependence of the observed responses of a material. In the simplest case, there is only a single characteristic time  $\tau$ , the origin of which goes back to Debye who proposed it in his seminal work on the dielectric response of polar liquids. If we define a shear compliance (the inverse of shear modulus)  $J(t)$  by

$$J(t) = \frac{\epsilon_{zx}}{\sigma_{zx}^0},$$

then, applying an oscillatory shear stress

$$\sigma_{zx}(t) = \sigma_{zx}^0 \exp(i\omega t),$$

on a polymer means imposing an oscillatory strain,  $\epsilon_{zx}(t) = \Delta J \sigma_{zx}(t)$ . Here,  $\Delta J$  is the *relaxation strength* of the material. The governing equation for  $\epsilon_{zx}(t)$  is then given by

$$\frac{d\epsilon_{zx}(t)}{dt} = -\frac{1}{\tau} [\epsilon_{zx}(t) - \Delta J \sigma_{zx}^0 \exp(i\omega t)]. \quad (22)$$

Assuming a solution,  $\epsilon_{zx}(t) = \sigma_{zx}^0 J^*(\omega) \exp(i\omega t)$ , where  $J^*(\omega)$  is the *complex shear compliance*, substituting it into Eq. (22) and solving it, yield

$$J^*(\omega) = \frac{\Delta J}{1 + i\omega\tau}, \quad (23)$$

which is usually referred to as a *Debye process*.

As discussed above, however, the dynamics of polymers and gels in the reaction bath cannot be described by a single relaxation time, rather by a statistical distribution of such characteristics times. For example, for the shear properties, we write the dynamic compliance  $J^*(\omega)$  as a sum of the Debye processes with relaxation times  $\tau_i$  and relaxation strengths  $\Delta J_i$ , so that

$$J^*(\omega) = J_u + \sum_i \frac{\Delta J_i}{1 + i\omega\tau_i}. \quad (24)$$

The sum is usually replaced by an integral, so that

$$J^*(\omega) = J_u + \int \frac{\mathcal{R}(\tau)}{1 + i\omega\tau} d\tau, \quad (25)$$

where  $\mathcal{R}(\tau)$  is called the *retardation time spectrum* of the shear compliance  $J^*$ . One may also write these results in terms of the complex modulus  $G_{\text{eff}}^*(\omega) = 1/J^*(\omega)$ , which then yields

$$G^*(\omega) = G_u - \int \frac{H(\tau)}{1 + i\omega\tau} d\tau, \quad (26)$$

where  $H(t)$  is the *relaxation time spectrum* of the complex modulus  $G^*(\omega)$  [or  $G(t)$ ]. In practice,  $H(t)$  cannot be directly measured, but is inferred only indirectly.

### Dynamic Mechanical Experiments

The evolution of the molecular structure of a polymer during the gelation process has a profound effect on the molecular mobility, which can be monitored by probing the changes in the viscosity and elastic moduli. The initial ( $p = 0$ ) liquid system has a steady shear viscosity  $\eta$  which increases with the extent of the reaction as the average molecular weight  $M_w$  increases. At the GP, the viscosity and the longest relaxation time  $\tau_{\text{max}}$  diverge. Beyond  $p_c$ , the equilibrium elastic moduli increase until they attain their highest values, which is when the reaction is brought to completion, i. e., when  $p \rightarrow 1$ .

All the experimental data for the elastic moduli of the nearly-critical gel network indicate that the effective elastic moduli  $G_{\text{eff}}$  follow a power law:

$$G_{\text{eff}} \sim (p - p_c)^z, \quad p > p_c. \quad (27)$$

On the other hand, near the GP, the viscosity of the sol phase also follows a power law, resulting in its divergence at the GP:

$$\eta \sim (p_c - p)^{-k}, \quad p < p_c, \quad (28)$$

while, as shown below, the longest relaxation time diverges as

$$\tau_{\text{max}} \sim |p - p_c|^{-k-z}, \quad |p - p_c| \ll 1. \quad (29)$$

The divergence of the viscosity at the GP is precisely due to the divergence of the mean polymer (cluster) size at the GP which, near the GP, follows a power law similar to Eq. (6) for the weight-average molecular weight  $M_w$  with precisely the same exponent  $\gamma$ .

In practice, it is precisely the divergence of  $\eta$  that signals the formation of the critical gel network. Due to the divergence of  $\tau_{\text{max}}$  measurements of the viscosity and elastic moduli fail at  $p_c$ , since steady-state conditions cannot be reached in a finite time. Another difficulty is that precise measurement of the GP is often difficult. Such difficulties

are partially overcome by performing *dynamic mechanical experiments*. In such experiments the sample is exposed to a periodically varying stress field. For example, a tensile stress  $\sigma_{zz}(t)$  is used

$$\sigma_{zz}(t) = \epsilon_{zz}^0 \exp(i\omega t), \quad (30)$$

which results in a time-dependent longitudinal strain  $\epsilon_{zz}(t)$  that varies with the frequency of the stress, but shows, in general, a phase-lag  $\varphi$ , such that

$$\epsilon_{zz}(t) = \epsilon_{zz}^0 \exp[-i(\omega t - \varphi)]. \quad (31)$$

We may then employ a *dynamic tensile modulus*  $G^*(\omega)$ , defined as

$$G^*(\omega) = \frac{\sigma_{zz}(t)}{\epsilon_{zz}(t)} = G'(\omega) + iG''(\omega). \quad (32)$$

Analogous experiments can, of course, be carried out for other types of mechanical loading. Of particular interest are measurements under simple shear which determine the relation between the shear strain  $\epsilon_{zx}$ , yielding the displacement along  $x$  per unit distance normal to the shear plane  $z = \text{constant}$ , and the shear stress  $\sigma_{zx}$  that acts on the shear plane along  $x$ .

In any case, such dynamic mechanical experiments measure the small amplitude oscillatory shear behavior of evolving gels. Under this condition, the gel evolution is continuous (no singularity). But, even such experiments cannot entirely overcome the difficulties in the determination of the exponent  $k$  of the viscosity, because the measurements cannot be carried out *at* the GP and in the limit of *zero frequency*. To estimate  $k$  one usually measures the frequency-dependent complex modulus  $G^*(\omega)$  at frequency  $\omega$ . At the GP and for low frequencies, one has

$$G' \sim G'' \sim \omega^n, \quad p = p_c, \quad (33)$$

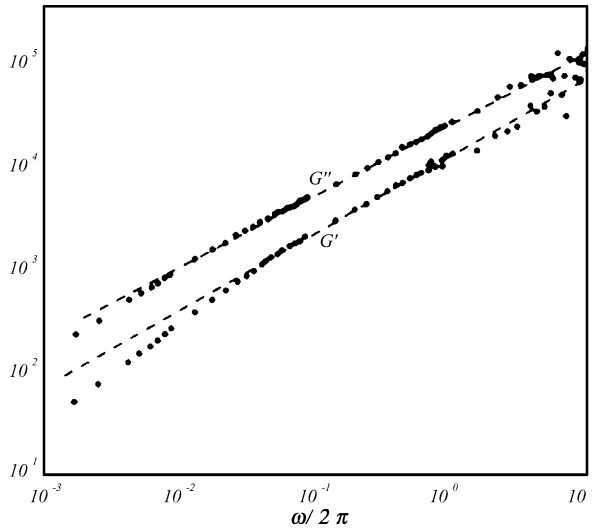
with

$$n = \frac{z}{z + k}, \quad (34)$$

where  $G'$  (the storage modulus) and  $G''$  (the loss modulus) describe storage and dissipation in an oscillating strain field of constant amplitude. Note that the exponent  $n$  in Eq. (34) is the same as that in Eq. (20). Typical variations of  $G'$  and  $G''$  with  $\omega$  are shown in Fig. 4 for a polycondensed gel very close to the GP.

The complex modulus  $G^*(\omega)$  is sometimes written as  $G^* = G + i\omega\eta$ , for which Durand et al. [29] proposed that

$$G^*(\omega, \epsilon) \sim \epsilon^z h_3(i\omega|p - p_c|^{z+k}), \quad (35)$$



**Percolation and Polymer Morphology and Rheology, Figure 4**  
Frequency-dependence of the storage modulus  $G'$  and loss modulus  $G''$  for a polycondensed gel close to the gel point (after Durand et al. [29])

where  $h_3(x)$  is a universal scaling function. The significance of scaling Eq. (35) is that it enables one to collapse the data for all values of  $|p - p_c|$  and  $\omega$  onto a single curve, usually called the *master curve* by polymer researchers. In the low-frequency regime, we do not expect  $G^*(\omega)$  to depend on  $|p - p_c|$ , but only on  $\omega$ , in which case one finds that,  $G^* \sim (i\omega)^n$ , which is equivalent to Eq. (33). Moreover, there is a loss angle  $\delta$  defined by,  $\tan \delta = G'/G''$ . The remarkable property of  $\delta$  is that at the GP it takes on a value  $\delta_c$  given by

$$\delta_c = \frac{\pi}{2}(1 - n) = \frac{\pi}{2} \frac{k}{z + k}, \quad (36)$$

so that, if the exponents  $z$  and  $k$  are universal, so will also be the loss angle  $\delta_c$ .

### Self-Similar Relaxation Time Spectrum

The observed power-law behavior of  $G'$  and  $G''$ , Eq. (33), implies a relaxation time spectrum which is self-similar (in time):

$$H(t) = \frac{G_0}{\Gamma(n)} \left( \frac{t}{\tau_0} \right)^{-n}, \quad (37)$$

where  $G_0$  is the characteristic modulus,  $\tau_0$  is the characteristic *shortest relaxation time*, and  $\Gamma$  is the gamma function. The modulus of a fully crosslinked polymer network is typically  $10^6$ – $10^7$  Pa, while the relaxation time of the network strand is about  $10^{-7}$ – $10^{-4}$  sec. The spectrum  $H(t)$

extends from the shortest time, at which the strands are beginning to be probed, to the infinite relaxation time of the critical gel network. The parameters  $G_0$  and  $\tau_0$  are material characteristics of the gel system. Most gel systems seem to possess the same value of  $n$ . However, there are also gels which exhibit no apparent universality in the value of  $n$ .

Viscosity and elastic moduli are rheological and mechanical properties of branched polymers and gels which characterize the dynamics of the polymerization, since we may measure indirectly the distribution of the relaxation times  $H(t)$  in the reaction bath. The moments of  $H(t)$  are directly related to the viscosity and the elastic moduli. Using Eqs. (35) and (37), we can back-calculate  $H(t)$  [23]:

$$H(t) \sim t^{-n} h_4(t|p - p_c|^{z+k}), \quad (38)$$

where  $h_4$  is another universal scaling function. Equation (38), which indicates that in the scaling regime near the GP the relaxation time distribution is a slowly decaying power law, generalizes Eq. (37) to any value of  $p$ , the extent of the polymerization. Equations (37) and (38) indicate that *any* relaxation property in the intermediate time or frequency range is *not* exponential, but follows a *power law*. As pointed out by Daoud [23], two distinct averages or characteristic times may be defined. One is

$$\tau_1 = \frac{\int H(t)dt}{\int [H(t)/t]dt} \sim |p - p_c|^{-k} \propto \eta, \quad (39)$$

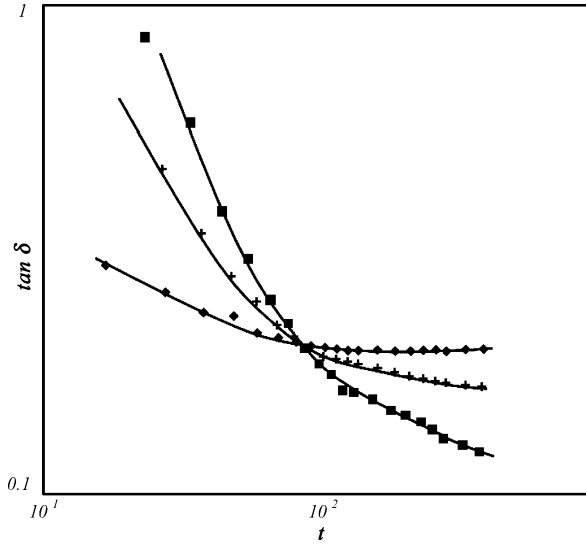
while the second one is given by

$$\tau_2 = \frac{\int tH(t)dt}{\int H(t)dt} \sim |p - p_c|^{-z-k}. \quad (40)$$

Note that  $\tau_2$  is in fact identical with  $\tau_{\max}$ , the longest relaxation time of the gel; see Eq. (29).

### Determination of the Gel Point

As already mentioned above, an important problem in polymerization and gelation is accurate determination of the GP, either for avoiding it in order to prevent gelation (so that a branched polymer with specific properties can be prepared), or for making polymeric materials very close to the GP, as they have unusual properties in that region. The GP, which is the analogue of a percolation threshold, depends on the functionality  $Z$  of the polymer – the analogue of the coordination number in the lattice models – and decreases with increasing  $Z$ . Thus, polymers with crosslinks of high functionality gel very early. Holly et al. [39] proposed using the loss angle  $\delta$  for locating the GP. They argued that, because as the GP is reached  $\tan \delta$  becomes independent of the frequency [see Eq. (36)], then, if one plots



**Percolation and Polymer Morphology and Rheology, Figure 5**  
Determination of gel point from data for loss angle  $\delta$ . Time is in minutes. The data are for frequencies  $31.6 \text{ rad s}^{-1}$  (diamonds),  $1.0 \text{ rad s}^{-1}$  (+) and  $0.0316 \text{ rad s}^{-1}$  (squares) (after Lin et al. [45])

$\tan \delta$  versus time at different frequencies, the intersection of all the curves should be at the GP. Figure 5 demonstrates how this method is used for locating the GP for a physical gel.

### Resistor and Elastic Percolation Networks

Experimental data for the scaling properties of the elasticity and viscosity of gels are usually compared with those of the conductivity and elasticity of percolation networks. Thus, we first briefly summarize percolation models of the conductivity of a percolation network. Consider a two-component network in which each (randomly-selected) bond has a conductance  $g_1$  with probability  $p$  or a conductance  $g_2$  with probability  $1 - p$ . The limit in which  $g_2 = 0$  and  $g_1$  is finite corresponds to a conductor-insulator mixture. As  $p \rightarrow p_c$ , more and more bonds are insulating, the conduction paths become very tortuous and, therefore, the effective conductivity  $g_{\text{eff}}$  of the network decreases. At  $p_c$  one has,  $g_{\text{eff}}(p_c) = 0$ , since no sample-spanning conduction path exists any more. In the critical region near  $p_c$  the effective conductivity follows a power laws:

$$g_{\text{eff}}(p) \sim (p - p_c)^t \quad \text{conductor-insulator networks.} \quad (41)$$

The limit in which  $g_1 = \infty$  and  $g_2$  is finite represents a *conductor-superconductor* mixture. All quantum-mechanical aspects of real superconductors are ignored in this definition, as we are concerned only with the effect of the local connectivity of the material on  $g_{\text{eff}}$ . It

should be clear that the effective conductivity  $g_{\text{eff}}$  of the network is dominated by the superconducting bonds. If  $p < p_c$ , then a sample-spanning cluster of the superconducting bonds does not exist, and  $\sigma_{\text{eff}}$  is finite. As  $p \rightarrow p_c^-$ ,  $g_{\text{eff}}$  increases until a sample-spanning cluster of the superconducting bonds is formed for the first time at  $p = p_c$ , where  $g_{\text{eff}}$  diverges. As  $p \rightarrow p_c^-$ , the effective conductivity follows a power law in the critical region:

$$g_{\text{eff}}(p) \sim (p_c - p)^{-s} \quad \text{conductor-superconductor networks.} \quad (42)$$

The exponents  $t$  and  $s$  are mostly universal. The article by Balberg discusses the conditions under which they may be non-universal.

In a similar manner, the elastic moduli of a two-phase percolation network are defined. Consider a two-component network in which each bond is an elastic element (a spring or beam) with an elastic constant  $e_1$  with probability  $p$  or  $e_2$  with probability  $1 - p$ . The limit in which  $e_2 = 0$  and  $e_1$  is finite corresponds to a mixture made of rigid materials and holes (for example, porous solids), or rigid and liquid materials. In such networks, as  $p \rightarrow p_c$ , an increasingly larger fraction of the bonds have no rigidity, the paths for transmission of stress or elastic forces become very tortuous and, therefore, the effective elastic moduli  $G_{\text{eff}}$  (Young's, bulk, or shear moduli) decrease. We refer to this model as the elastic percolation network. At  $p_c$  one has,  $G_{\text{eff}}(p_c) = 0$ , while near  $p_c$  in the critical region,

$$G_{\text{eff}}(p) \sim (p - p_c)^T \quad \text{rigid-soft two-phase networks.} \quad (43)$$

The limit in which  $e_1 = \infty$  and  $e_2$  is finite represents mixtures of rigid-superrigid materials. We refer to the model as the superelastic percolation network. In this case the effective elastic moduli  $G_{\text{eff}}$  of the network are dominated by the superrigid bonds. If  $p < p_c$ , then a sample-spanning cluster of the superrigid bonds cannot form, and  $G_{\text{eff}}$  is finite. As  $p \rightarrow p_c^-$ , the effective elastic moduli increase until the percolation threshold  $p_c$  of the rigid phase is reached, at which a sample-spanning cluster of the superrigid bonds is formed for the first time, and the effective elastic moduli diverge. In the critical region near  $p_c$  one has

$$G_{\text{eff}}(p) \sim (p_c - p)^{-S} \quad \text{rigid-superrigid networks.} \quad (44)$$

Unlike the exponents  $t$  and  $s$  for resistors networks, the exponents  $T$  and  $S$  may depend on the details and the type of the forces that are active in the networks. Thus, in what

follows, we first briefly describe various types of forces that have been included in such networks.

### The Born Model

The Born model [17] is described by the following elastic Hamiltonian,

$$\mathcal{H} = \frac{1 + \mathcal{P}}{1 - \mathcal{P}} \sum_{ij} \mu [(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{R}_{ij}]^2 + \frac{1 - 3\mathcal{P}}{4(1 - \mathcal{P})} \sum_{ij} (\mathbf{u}_i - \mathbf{u}_j)^2, \quad (45)$$

where  $\mathcal{P}$  is the Poisson's ratio,  $\mu$  the shear modulus,  $\mathbf{u}_i$  the displacement of site  $i$ , and  $\mathbf{R}_{ij}$  the unit vector along the line (lattice bond) that connects sites  $i$  and  $j$ . The first term of Eq. (45) is the energy of a network of central-force (CF) springs, i. e., Hookean springs that transmit force only in the  $\mathbf{R}_{ij}$  direction, but do not transmit shear forces. The second term is a contribution analogous to, for example, the power dissipated in conduction, since  $(\mathbf{u}_i - \mathbf{u}_j)^2$  represents the *magnitude* of the displacement difference  $\mathbf{u}_i - \mathbf{u}_j$ . The Born model may be considered as an analogue of a 3D solid in plane-stress with holes normal to the  $x - y$  plane, or as a 2D solid with the Poisson's ratio defined as the negative of ratio of the strain in the  $y$ -direction to that in the  $x$ -direction, when a stress is applied in the  $x$ -direction but none is applied in the  $y$ -direction. Results for a 3D solid in plane-strain can be generated from those of this model using the transformation,  $\mathcal{P}' = \mathcal{P}/(1 + \mathcal{P})$ , where  $\mathcal{P}'$  is the Poisson's ratio for the plain strain.

The Born model does suffer from some peculiarities. For example, it is not difficult to show (although it may not be obvious at first glance) that, except for  $\mathcal{P} = 1/3$ , the elastic energy  $\mathcal{H}$  defined by Eq. (45) is *not* invariant with respect to arbitrary rigid body rotations, a fundamental requirement for any reasonable model of elastic properties of materials. In the limit  $\mathcal{P} = 1/3$  the model reduces to a network of CF springs. When the elastic energy of a system is written in terms of an expansion in the displacement field  $\mathbf{u}$ , its rotational invariance is not easy to see. To demonstrate the lack of rotational invariance of the elastic energy, one substitutes an infinitesimal rotation  $\boldsymbol{\omega} \times \mathbf{R}_i$  for the displacement vector  $\mathbf{u}_i$ , where  $\mathbf{R}_i$  is the position vector of  $i$ . An arbitrary rotation of the solid should not contribute to its energy, but Eq. (45) indicates that, while the contribution of the CF part would indeed be zero, that of the scalar-like part would not be and, therefore,  $\mathcal{H}$  is not rotationally invariant. Moreover, although materials do exist that have a Poisson's ratio as high as  $1/2$  ( $\mathcal{P}$  can theoretically

be as high as 1 in 2D materials [58]), the model fails to have a strictly positive energy for  $\mathcal{P} > 1/3$  and, therefore, violates the thermodynamic requirement that the potential energy be a minimum at zero strain. Another example of displacements that contribute to the scalar-like portion of the energy  $\mathcal{H}$  of the model, but not to the CF portion, arises when a significant fraction of the lattice's bonds is removed, i. e., a percolation network is generated. In such a lattice a site that is connected to only one bond can have an arbitrary displacement in the direction orthogonal to the direction of the bond without affecting the CF part of the elastic energy, as can a site which is connected to only a set of two collinear bonds.

In his original formulation of this model, Born [17] inserted the scalar-like part of the elastic energy (45) as a substitute for the many-body, angular and bending terms (see below) that normally arise in describing the elastic properties of materials, because the expansion of such scalar two-body terms is much simpler and more convenient than expanding the many-body terms that they replace. When viewed in this way, the coefficients of the model should be treated as fitting parameters. Hence, let us rewrite Eq. (45) as

$$\mathcal{H} = \frac{1}{2}\alpha_1 \sum_{ij} [(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{R}_{ij}]^2 + \frac{1}{2}\alpha_2 \sum_{ij} (\mathbf{u}_i - \mathbf{u}_j)^2, \quad (46)$$

where  $\alpha_1$  and  $\alpha_2$  now represent two adjustable parameters. When introduced in this context, one may use the Born model for modeling and fitting elastic properties of certain materials.

Note that, so long as  $\alpha_2 > 0$ , the scalar-like term of Eq. (45) or (46) is the dominating contributing factor to the elastic energy  $\mathcal{H}$ . This implies immediately that, although the Born model is a vector model, the behavior of the elastic moduli in this model near the percolation threshold is effectively like that of a scalar (conductivity) model and, therefore,

(i) the percolation threshold of the Born model, at which the elastic moduli vanish or diverge, is the same as that of random percolation models, and

(ii) near the percolation threshold the elastic moduli of the Born model follow power law (43) or (44), but with  $T = t$  and  $S = s$ . That is, the power-law behavior of the elastic moduli in the Born model is the same as that of the effective conductivity.

### The Central-Force Model

Consider the limit  $\mathcal{P} = 1/3$  of Eq. (45), i. e., a network of CF or Hookean springs. Since the elastic materials that we wish to consider are heterogeneous, the local shear modu-



lus  $\mu$  varies spatially. Thus, writing  $\mu = e_{ij}\alpha/4$  and taking  $P = 1/3$  reduce Eq. (45) to

$$\mathcal{H} = \frac{1}{2}\alpha \sum_{ij} [(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{R}_{ij}]^2 e_{ij} \quad (47)$$

where  $\alpha$  is the CF constant.

We can easily compute the elastic moduli of the CF networks, if no percolation effect is present, i. e., if no bond is broken and all the  $e_{ij}$  are equal. Suppose that each spring has an unstretched length  $\ell_0$ . Then, it is not difficult to show that the bulk modulus  $K$  of a triangular network is given by [58]

$$K = \frac{\sqrt{3}}{2}\alpha \quad \text{triangular network,} \quad (48)$$

whereas its shear modulus is given by [58]

$$\mu = \frac{\sqrt{3}}{4}\alpha \quad \text{triangular network,} \quad (49)$$

and, therefore, the Poisson's ratio of the network is,  $P = (K_e - \mu_e)/(K_e + \mu_e) = 1/3$ . Similarly, we may show that [58],

$$K = \mu_p = \frac{1}{2}\alpha \quad \text{square network,} \quad (50)$$

where  $\mu_p$  is the shear modulus in *pure* shear (the network's shear modulus in *simple* shear is zero). As for the standard 3D cubic networks, one has [58]

$$K_e = \begin{cases} \frac{1}{3\ell_0}\alpha & \text{simple-cubic network,} \\ \frac{1}{\sqrt{3}\ell_0}\alpha & \text{BCC network,} \\ \frac{2\sqrt{2}}{\sqrt{3}\ell_0}\alpha & \text{FCC network.} \end{cases} \quad (51)$$

A simple-cubic network does not possess a shear modulus in simple shear. We emphasize that Eqs. (48)–(51) are valid at zero temperature and when the external stress is infinitesimally small.

In practice, however, all the experimental measurements are carried out at temperatures above  $T = 0$  and, therefore, it is important to understand the temperature-dependence of the elastic moduli, at least in the context of the network models. In addition, in many practical situations, the material under study is exposed to a finite stress or tension (as opposed to an infinitesimal stress or tension considered above) and, thus, the role of such an external driving force in determining the elastic properties of materials must be understood. In principle, the role of the temperature can be understood by carrying out molecular dynamics simulations [54]. However, phenomenological calculations of the type described above can also be carried

out for homogeneous networks at non-zero temperature. Suppose, then, that a 2D isotropic tension  $\sigma_i$  is imposed on a network at a non-zero temperature. One can show that for the triangular network [58]

$$K = \frac{1}{2}(\sqrt{3}\alpha - \sigma_i), \quad (52)$$

$$\mu = \frac{\sqrt{3}}{4}(\alpha + \sqrt{3}\sigma_i), \quad (53)$$

which reduce to Eqs. (48) and (49) in the limit  $\sigma_i = 0$ . Similar results are obtained for the square network [58]:

$$K = \frac{1}{2}(\alpha - \sigma_i), \quad (54)$$

$$\mu_p = \frac{1}{2}(\alpha + \sigma_i), \quad (55)$$

$$\mu_s = \sigma_i, \quad (56)$$

where  $\mu_s$  is the shear modulus of the network in simple shear.

### Rigidity Percolation

If the elastic constants  $e_{ij}$  of the bonds of a CF network take on either a finite value with probability  $p$  or vanish with probability  $1 - p$ , then one obtains an elastic percolation network. If  $e_{ij}$  is infinitely large with probability  $p$  or takes on a finite value with probability  $1 - p$ , then, one obtains a superelastic percolation network. Percolation on such networks of Hookean springs is called the *rigidity percolation*. Such networks are of both theoretical and practical interest. In addition to the polymeric materials described in this chapter, they are also useful models for describing the elastic properties of biological materials. Moreover, in many engineering problems, structures composed of bars or beams connected at nodes that are called trusses acquire their rigidity mainly from the tensile and compressive stiffness of the beams, and these are CF type of contributions. For example, in the absence of friction between the particles of a granular packing, which is a reasonable model of unconsolidated porous materials (such as powders), the mechanical behavior of the packing is similar to those of rigidity percolation. In contrast, those in which angular forces, e. g., covalent bonds at the molecular level, are the most important are usually referred to as *frames*. It is not difficult to see that rigid systems in which angular forces dominate their behavior support an applied stress, so long as they are simply connected. In contrast, the CF systems require higher degrees of connectivity. Therefore, the percolation thresholds of CF networks are much larger than those of random percolation networks; see the articles by Duxbury and Thorpe.

### The Bond-Bending Model

Consider an elastic percolation network in which there are both central and bond-bending (angle-changing) forces, with the latter type representing the three-body interactions. One of the main advantages of such a model is that their percolation threshold can be the same as that of random percolation, if the many-body interactions are such that any deformation of the lattice is done at some costs to its elastic energy. In general, the elastic energy of such models is given by [40]

$$\mathcal{H} = \frac{1}{2}\alpha \sum_{\langle ij \rangle} [(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{R}_{ij}]^2 e_{ij} + \frac{1}{2}\chi \sum_{\langle jik \rangle} (\delta\theta_{jik})^2 e_{ij} e_{ik}, \quad (57)$$

where  $\alpha$  and  $\chi$  are, respectively, the central and bond-bending (BB) force constants. Here,  $\langle jik \rangle$  indicates that the sum is over all triplets in which the bonds  $j-i$  and  $i-k$  form an angle with its vertex at  $i$ . The first term on the right-hand side of Eq. (60) represents the usual CF contributions (see above), while the second term is due to the BB (angle-changing) forces. The precise form of  $\delta\theta_{jik}$  depends on the microscopic details of the model. In the version that is of interest to us, bending of the collinear bonds is allowed, in which case [13,66]

$$\delta\theta_{jik} = \begin{cases} (\mathbf{u}_{ij} \times \mathbf{R}_{ij} - \mathbf{u}_{ik} \times \mathbf{R}_{ik}) \cdot \frac{(\mathbf{R}_{ij} \times \mathbf{R}_{ik})}{|\mathbf{R}_{ij} \times \mathbf{R}_{ik}|}, & \mathbf{R}_{ij} \text{ not parallel to } \mathbf{R}_{ik}, \\ |(\mathbf{u}_{ij} + \mathbf{u}_{ik}) \times \mathbf{R}_{ij}|, & \mathbf{R}_{ij} \text{ parallel to } \mathbf{R}_{ik}, \end{cases} \quad (58)$$

where,  $\mathbf{u}_{ij} = \mathbf{u}_i - \mathbf{u}_j$ . For all 2D networks, Eq. (61) is simplified to

$$\delta\theta_{jik} = (\mathbf{u}_i - \mathbf{u}_j) \times \mathbf{R}_{ij} - (\mathbf{u}_i - \mathbf{u}_k) \times \mathbf{R}_{ik}. \quad (59)$$

We refer to the model described by Eqs. (60)–(62) as the BB model. For most materials to which the BB model is applicable, one has,  $\chi/\alpha \leq 0.3$  [49]. Sahimi [57] suggested that the critical exponent  $T$  of the elasticity in the BB model is related to  $t$ , the critical exponent of conductivity of percolation networks:

$$T = t + 2\nu, \quad (60)$$

where  $\nu$  is the correlation length exponent of percolation. This relation is in excellent agreement with the available numerical estimates (see Table 1 below). The articles by Duxbury and Thorpe provide much more detailed discussions of the CF and BB models and, therefore, we do not discuss them any further.

Before embarking on a comparison between the experimental data for the rheological properties of gels near the GP, for convenience and as a basis for comparison with the experimental data, we summarize in Table 1 the current most accurate estimates of the various critical exponents for the conductivity and elasticity of percolation models near the percolation thresholds, including the CF and BB models.

We are now in a position to compare the predictions of the percolation models with the experimental data for the viscosity of the nearly critical sol, and the elastic moduli of the nearly critical gels.

### Nearly-Critical Chemical Gels: Comparison of the Data with the Percolation Models

There are numerous experimental measurements of the elastic moduli of nearly-critical chemical gels and the associated exponent  $z$ . Examples include the measurements for hydrolyzed polyacrylamide [9,11], tetraethylorthosilicate reactions [38,64], gelatin solutions [28], polycondensation of polyoxypropylated trimethylolpropane with hexamethylenediisocyanate [29], and several other measurements [2,30,35,37,65]. These measurements yielded esti-

#### Percolation and Polymer Morphology and Rheology, Table 1

Estimates of the critical exponents of the conductivity and elastic moduli of percolation models in  $d$ -dimensions. Values of  $T$  and  $S$  for the CF model refer to bond percolation, whereas those of  $t$  and  $s$ , the conductivity exponents, are independent of the model.  $\nu$  is the critical exponent of the percolation correlation length. Value of  $\nu$  for the CF model is different from that of random percolation, whereas it is the same as that of random percolation for the BB model

$d$	$t/\nu$	$s/\nu$	$T/\nu$	$S/\nu$	Model
2	$0.9745 \pm 0.0015$	$0.9745 \pm 0.0015$	$2.97 \pm 0.03$	$0.92 \pm 0.03$	Bond bending
	—	—	$2.95 \pm 0.25$	$0.92 \pm 0.02$	Central force
3	$2.27 \pm 0.01$	$0.835 \pm 0.005$	$4.3 \pm 0.1$	$0.74 \pm 0.04$	Bond bending
	—	—	$2.1 \pm 0.1$	$0.80 \pm 0.03$	Central force

mates of  $z$  that are in the range 1.9–2.4, which do not agree with the value of the critical exponent  $T$  for the 3D BB model (see Table 1). In fact, if the size of a chemical gel network is large enough, the BB forces may not play any important role in determining the elastic properties of nearly critical gels and, therefore, the only important forces between the monomers are the central (stretching) forces. Therefore, these experimental data may be explained based on the elasticity exponent in the 3D CF percolation [14],  $T \simeq 2.1 \pm 0.2$ .

However, a value of  $z$  in the range 1.9–2.4 may also be interpreted in terms of two other models:

(i) As mentioned earlier, de Gennes [72] suggested that the scaling properties of the elastic moduli of nearly critical gels are in the universality class of the conductivity of percolation networks, implying that,  $z = t \simeq 2.0$ .

(ii) On the other hand, Alexander [7] argued that in some gels and rubbers that are under internal or external stresses, there are terms in their elastic energy that are similar to the Born model. As discussed above, the critical exponent of the elastic moduli in the Born model is equal to that of the conductivity,  $t$ , and in particular in 3D,  $T = t \simeq 2.0$ , because near the percolation threshold the contribution of the second term of the right side of Eq. (45) or (46), which is a purely scalar term, dominates that of the first term which is due to the CFs.

While the data mentioned above are more or less consistent with de Gennes' hypothesis, most of them are not precise enough to distinguish between  $t \simeq 2.0$  for 3D percolation conductivity and  $T \simeq 2.1$  for the CF percolation. However, there are also a few relatively precise sets of experimental data that seem to support de Gennes' conjecture. For example, Axelos and Kolb [15] measured the rheological properties of pectin biopolymers that consist of randomly connected  $\alpha(1-4)$ D-galacturonic acid units and their methyl esters. If the methyl ester content is low, pectin forms thermoreversible gels upon addition of cations, such as calcium. Axelos and Kolb [15] measured the frequency dependence of the storage modulus  $G'(\omega)$  and loss modulus  $G''(\omega)$  [see Eq. (33)] and reported that,  $z \simeq 1.93$ ,  $k \simeq 0.82$ , and  $n \simeq 0.71$ . Their elasticity exponent is close to that of the conductivity,  $t \simeq 2.0$ , for 3D resistor networks. Less precise data, but still supportive of de Gennes' proposal, were reported by Adam et al. [5] for the complex modulus of end-linked poly(dimethylsiloxane) pregel polymer clusters, quenched at different distances from the gelation threshold. They reported that,  $z \simeq 1.9 \pm 0.15$ , consistent with the value of the conductivity exponent  $t$  for 3D percolation. However, the estimated error is large enough that one can easily in-

terpret such a value of  $z$  in terms of the CF percolation model as well.

At first glance, de Gennes' proposal that the critical exponent of the effective moduli of gels, a vector transport property, should be equal to that of a scalar property, the effective electrical conductivity of a resistor network, may seem incorrect. However, to justify his proposal, de Gennes introduced the notion of an elastic chain between neighboring nodes or monomers that are the analogue of quasi-one-dimensional strands that percolation clusters possess near the percolation threshold (see the article by Coniglio). He then argued that if such chains are elongated, then their nodes carry an extra amount of energy. If we assume that the blobs – the multiply-connected parts – of the large cluster of monomers do not contribute significantly to the elastic moduli, then one must only consider the energetics of the links or the chains. If the extra energy of such chains is larger than  $k_B T$  ( $k_B$  is the Boltzmann's constant), then, as Daoud [24] argued, one obtains de Gennes' proposal [72],  $t = z$ , although Daoud's analysis was a mean-field approximation, not a scaling one.

As for Alexander's proposal [7], rubbers and gels differ from the Born model in several important ways, such as the presence of non-linear terms in their elastic energy, and the possibility of negative as well as positive Born coefficients  $\alpha_1$  and  $\alpha_2$  in Eq. (46). Therefore, as discussed above, while one may use the Born model to *fit* the experimental data, it is not clear that, at a fundamental level, the Born model can actually describe the elastic properties of such gels, since its elastic energy is not rotationally invariant.

### Enthalpic Versus Entropic Elasticity

There is yet another way of rationalizing the experimental data for the scaling properties of nearly critical gels. To describe it, we mention that several measurements of the elastic moduli of nearly critical gel and the associated exponent  $z$  deviate significantly from all the data described above. Examples include the measurements of Adam et al. [1] for polycondensation,  $z \simeq 3.3 \pm 0.5$ , those of Martin et al. [47] and Adolf et al. [6] for gels made from 89% (by weight) of the diglycidyl ether of bisphenol A cured with 11% (by weight) of diethanolamine which yielded,  $z \simeq 3.3 \pm 0.3$ , and the data reported by Colby et al. [22] for polyester gels, which have been argued to lie in the middle of the static crossover between the Flory–Stockmayer (Bethe lattice) model, which predicts that,  $z = 3$ , and the 3D percolation model. Colby et al. [22] reported that,  $z \simeq 3.0 \pm 0.7$ , which is inconsistent with both the CF percolation and the BB models, although one

might argue that their estimate is agreement with  $z = 3$ , the Flory–Stockmayer value of  $z$ . More recent measurements of the shear modulus of an end-linking polymer gel network by Takahashi et al. [64] yielded,  $z \simeq 2.7$ , which is again in the range of the above data.

One possible explanation for such data is that the elasticity of such gels is entropic rather than enthalpic. Plischke and Joós [54], Plischke et al. [55], Farago and Kantor [32], and Plischke [52,53] argued that the CF and the BB models, described above, are applicable to gels at temperature  $T = 0$ , and that for  $T \neq 0$  there is an important contribution to the shear modulus that is entropic in nature. In analogy with the physics of rubber elasticity, Plischke et al. [55] argued that, near the percolation threshold or the GP, the polymer network consists essentially of long chains of singly-connected particles (monomers or sites), linked to each other at various junction points. Such chains are similar to the polymer chains that are crosslinked in rubber in order to produce a rigid amorphous material. Deformation of the sample changes the distance between the junction points or crosslinks, as a result of which the entropy is generically decreased, resulting in an increase of free energy and, hence, a restoring force. There would be a net shear-restoring force, when the nearly critical gel is formed, implying that the connecting chain of particles acts as a stretched spring. Molecular dynamics simulations (for a review see, for example, [42]) of Plischke and co-workers, and computer simulations of Farago and Kantor [32], who used a model that consisted of hard spheres in which a fraction  $p$  of the neighbors are tethered by inextensible bonds, both yielded,  $z \simeq t$ .

On the other hand, del Gado et al. [74] proposed a different model, also purported to be appropriate for entropic gels, in which one begins with a random collection of monomers with concentration  $p$ . Each pair of the monomers are then linked with a probability  $p_b$  to form permanent bonds. Varying  $p_b$  produces a distribution of clusters of the linked monomers and, hence, gives rise to the possibility of forming a sample-spanning cluster. Monomers and the clusters then diffuse according to the bond fluctuation algorithm of Carmesin and Kremer [21]. In this algorithm, the monomers diffuse in the solution randomly but obeying the excluded-volume interaction (i.e., no two monomers can occupy the same point in space). Due to this random motion, the bonds may have to change their length in a set of allowed values and, thus, they may have to bend and take on many different values of the angles between the bonds, which then gives rise to a wide variety of polymer conformations. The mean-square displacement  $\langle R^2(t) \rangle$  of the polymer's center of mass is then calculated which, due to the elastic potential

that reduces the fluctuations proportionally to the effective elastic constant  $\alpha$ , is given by

$$\langle R^2(t) \rangle \propto \alpha^{-1}, \quad (61)$$

and, therefore, the elastic constant and its power-law behavior near the percolation threshold  $p_c$  can be computed, from which the exponent  $z$  is extracted. Two-dimensional simulations of del Gado et al. [74] yielded the estimate,  $z \simeq 2.7 \pm 0.1$ , hence disagreeing with the results of Plischke and co-workers, and Farago and Kantor.

How can one interpret these results? If the elasticity of these gels is due to entropic effects, then, as Daoud and Coniglio [25] argued (see also Martin et al. [47]), the elastic free energy  $\mathcal{H}$  per unit volume must be given by

$$\mathcal{H} \sim \xi_p^{-d} \alpha_\ell \xi_p^2, \quad (62)$$

where  $\xi_p$  is the correlation length of percolation, and  $\alpha_\ell$  is the effective elastic constant of a long chain of length  $\xi_p$  connecting two nodes. Since  $\alpha_\ell \sim \xi_p^{-2}$ , we obtain

$$z = \nu d. \quad (63)$$

Since for 2D percolation,  $\nu = 4/3$ , Eq. (66) predicts that,  $z = 8/3 \simeq 2.66$ , quite different from the exponents  $T$  of the CF and BB models, and also the conductivity exponent,  $t \simeq 1.3$  (see Table 1). However, it is in agreement with the numerical simulations of del Gado et al. [74]. Daoud [24] argued that Eq. (66) is valid when the energy of the chains is of the same order of magnitude as the thermal energy  $k_B T$ .

Since,  $\nu \simeq 0.89$  for 3D percolation, Eq. (66) predicts that,  $z \simeq 2.67$ , consistent with the experimental data of Adam et al. [1], Martin et al. [47], Adolf et al. [6], Colby et al. [22] and Takahashi et al. [64] mentioned above, all of whom reported estimates of the elasticity exponent  $z$  in the range  $2.7\text{--}3.3 \pm 0.5$ . Therefore, while these experimental data may be explained by Eq. (66), the numerical results of Plischke and co-workers [52,53,54,55], as well as those of Farago and Kantor [32], do not agree with the prediction of Eq. (66). Indeed, although the main argument of Plischke et al. [55] was that, the entropic effects are important at temperatures  $T \neq 0$ , where one should see a crossover to the value of the conductivity exponent, all of the above experiments were also carried out at finite temperatures, yet they did not indicate that,  $z \simeq t$ .

More recently, Xing et al. [69] studied the scaling of shear modulus near the gelation-vulcanization transition. They proposed that in a dense melt the sizes of the effective chains of the critical gel scale *sublinearly* with their counter length. The implication is that, the energy

that each chain contributes is of the order of  $k_B T$ , hence leading to Eq. (66). However, in *phantom* networks – those in which there is no repulsion between the particles (monomers) – the chains' sizes scale *linearly* with their contour length, which means that the elasticity exponent  $z$  crosses over to the conductivity exponent  $t$ . Thus, it may be that some of the conflicts between the various experimental estimates of  $z$  are due to the crossover effects, but the issue remains unsolved.

### Viscosity of Nearly Critical Sol: Comparison of the Data with the Percolation Models

There is also a wealth of experimental data for the viscosity of the nearly critical sol solution and the associated critical exponent  $k$ , defined by Eq. (28). An important question that has been studied for over two decades is: how can one explain the extensive experimental data for the scaling behavior of the viscosity  $\eta$  of the sol phase near the GP? To begin with, it was proposed by Sahimi and Goddard [60] (see also [12,59]) that the power-law behavior of  $\eta$  near the GP is analogous to that of the shear modulus of a superelastic percolation network near  $p_c$  and, therefore, the same critical exponent characterizes both. To proceed further, we must first establish a rigorous relationship between the linear elasticity and the theory of viscous fluids, thus confirming the proposal of Sahimi and Goddard [60].

We consider a general time-dependent system and write down the equation of motion for a macroscopically-homogeneous material in terms of the displacement field  $\mathbf{u}$ :

$$(\lambda + \mu)\nabla(\nabla \cdot \mathbf{u}) + \mu\nabla^2 \mathbf{u} + \mathbf{F} = \rho \frac{\partial^2 \mathbf{u}}{\partial t^2}, \quad (64)$$

where  $\rho$  is the mass density,  $t$  the time,  $\lambda$  and  $\mu$  the usual Lamé coefficients, and  $\mathbf{F}$  an external force. For an *incompressible* material, i. e., one for which the bulk modulus  $K$  and the Lamé coefficient  $\lambda$  are both divergent, one has the solenoidal condition,

$$\nabla \cdot \mathbf{u} = 0. \quad (65)$$

Due to the incompressibility condition, the first term of Eq. (67) is indeterminate. Equation (67) can be then written in terms of the reactive hydrostatic pressure  $P$ ,

$$-\nabla P + \mu\nabla^2 \mathbf{u} + \mathbf{F} = \rho \frac{\partial^2 \mathbf{u}}{\partial t^2}. \quad (66)$$

On the other hand, let us write down the Navier–Stokes equations of motion for an incompressible and Newtonian viscous fluid,

$$-\nabla P + \eta\nabla^2 \mathbf{v} + \mathbf{F} = \rho \left( \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right), \quad (67)$$

where  $\mathbf{v}$  is the fluid velocity field, and  $\eta$  the fluid's dynamic viscosity. For an incompressible fluid, the continuity equation is given by

$$\nabla \cdot \mathbf{v} = 0, \quad (68)$$

which is similar to Eq. (68). For slow fluid flow, i. e., when the Reynolds number  $\text{Re} \ll 1$ , the non-linear inertial term,  $\mathbf{v} \cdot \nabla \mathbf{v}$ , is very small and can be neglected, which means that the Navier–Stokes equations reduce to

$$-\nabla P + \eta\nabla^2 \mathbf{v} + \mathbf{F} = \rho \frac{\partial \mathbf{v}}{\partial t}. \quad (69)$$

Thus, we see that, under steady-state condition, and when the flow of the fluid is slow, the governing equations for the displacement field  $\mathbf{u}$  and the velocity field,  $\mathbf{v} = \partial \mathbf{u} / \partial t$ , are exactly identical, provided that there is a one-to-one correspondence between the shear modulus  $\mu$  and the dynamic viscosity  $\eta$ .

In addition, under certain conditions, the effective viscosity  $\eta$  of a suspension of completely rigid spheres in creeping (very slow) flow of an incompressible fluid of viscosity  $\eta_1$  is related to the *steady-state* effective shear modulus  $\mu$  of a two-phase material composed of the same completely rigid spheres in an incompressible matrix with shear modulus  $\mu_1$ . In this case, the working equation is given by

$$\frac{\eta_e}{\eta_1} = \frac{\mu_e}{\mu_1}. \quad (70)$$

Equation (70) is exact when, regardless of the configuration of the particles, hydrodynamic interactions between the particles can be neglected, which is the case when the system is infinitely dilute so that the volume fraction of the particles approaches zero. Even if the system is non-dilute, Eq. (70) would still be exact, provided that the configurations of the particles in the flow and the elasticity problems are identical.

Having established a theoretical connection between the viscosity  $\eta$  of a sol and the shear modulus of an appropriate two-phase material, the one-to-one correspondence between  $\eta$  and the shear modulus of a superelastic percolation network should be clear because,

- (i)  $\eta$  and  $\mu$  both diverge at  $p_c$  (the GP), and
- (ii) the percolation models predict accurately the structure and elastic moduli of nearly critical gels.

On the other hand, de Gennes [73] (see also Alain et al. [8]) suggested an analogy between  $\eta$  and the effective conductivity of a conductor-superconductor percolation networks, the effective conductivity of which diverges at  $p_c$  according to power law (42), so that  $k = s$ .



However, even a one-to-one correspondence between  $\eta$  and the shear modulus of a superelastic percolation network is not nearly enough to explain the power law behavior of the viscosity of gelling solutions near the GP, because most of the experimental data indicate that the value of  $k$  is either in the range 0.6–0.9 (see, for example, [1,4,10,29]), or in the range 1.3–1.5 (see, for example, [28,47,48]), whereas the power-law behavior of the shear modulus of a 3D superelastic percolation network near  $p_c$  is characterized by a *unique* value of the exponent  $S$  defined by Eq. (44). The reason for having two distinct values of  $k$  is [12,59] that the dynamics of the sol solutions, that yield the two distinct values of  $k$ , may be completely different.

In one case, the solution may be close to the Zimm limit in which there is little or no polymer diffusion in the reaction bath, because there are strong hydrodynamic interactions between the monomers, and also between polymers of various sizes, that prevent diffusion in a nearly critical sol near the GP. Hence, a superelastic percolation network – a *static* system with fixed rigid clusters – may be suitable for simulating the Zimm regime. For this limit, Arbabi and Sahimi [12] suggested that,

$$S = \nu - \frac{1}{2}\beta, \quad (71)$$

where  $\nu$  and  $\beta$  are the standard percolation exponents. Then, with  $\nu \simeq 0.89$  and  $\beta \simeq 0.41$  for 3D percolation, one obtains,  $S \simeq 0.68$  for the divergence of the shear modulus of 3D superelastic percolation networks at the percolation threshold. Such a value of  $S$  seems to explain the estimates of  $k$  for those gelling solutions that have a viscosity exponent in the range 0.6–0.9, implying that the scaling behavior of their viscosity near the GP may be described by the shear modulus of a *static* superelastic percolation network.

On the other hand, the gelling solution may also be in, or near, the Rouse regime – one in which there are no hydrodynamic interactions between the polymers of various sizes – and, therefore, the finite polymers can diffuse essentially freely in the reaction bath. To simulate this regime the following model was proposed [12,59]. We consider a superelastic percolation network in which every cluster of the totally rigid bonds (the bonds in such clusters are totally rigid in order to distinguish them from those with a finite elastic constant) represents a finite polymer. Due to randomness of percolation networks, there is, of course, a wide distribution of such polymers or clusters in the network. The “soft” bonds – those with a finite elastic constant – represent the liquid solution in which the rigid clusters move randomly, with equal probability, in any direction of the network. The motion simulates diffusion of the finite polymers in the reaction bath. Two rigid clusters

cannot overlap, but can temporarily join and form a larger cluster, which may also be broken up again at a later time. Thus, at every time step, a cluster and a direction for the motion are picked at random, and the cluster is moved by one step (one lattice bond) in that direction, taking into account the excluded-volume effect. This is then a *dynamic* superelastic percolation network, the shear modulus of which can be calculated at long times. Monte Carlo simulations [12,59] indicated that the shear modulus of such a dynamic superelastic percolation network diverges with an exponent  $S'$ , which Arbabi and Sahimi [12] proposed it to be given by

$$S' = 2\nu - \beta. \quad (72)$$

Equation (72) predicts that in 3D,  $S' \simeq 1.35$ , which appears to explain the viscosity exponent for those gelling solutions that have an exponent  $k$  in the range 1.3–1.5, hence implying that the scaling behavior of their viscosity near the GP may be described by the shear modulus of a *dynamic* superelastic percolation network.

Daoud [24] also argued that, similar to the case of the elastic moduli of nearly critical gels described above, one must consider two distinct regimes for explaining the power-law behavior of the viscosity near the GP, except that his arguments were based on the energetics of the solution. According to him, if the elastic chains carry an energy which is of the same order of magnitude as the thermal energy  $k_B T$ , then the exponent  $S$  should be given by Eq. (71). On the other hand, if the elastic chains are stretched and have an extra energy larger than  $k_B T$ , then one should recover Eq. (72), which had also been conjectured by de Gennes [73], but based on the analogy between the viscosity and the effective conductivity of superconducting percolation networks described above.

We should point out that, as in the case of the elastic moduli of nearly critical gels discussed earlier, there are some experimental data that indicate some deviations of  $k$  from  $S$  or  $S'$ . However, as pointed out earlier, experimental determination of  $k$  (and the elasticity exponent  $z$ ) involves (see, for example, [29]) measuring the complex modulus  $G^*(\omega)$  for a series of frequencies  $\omega$ . But, strictly speaking, the power laws for the elastic moduli of the elastic and superelastic percolation networks are valid only in the limit,  $\omega \rightarrow 0$ , whereas in practice it is essentially impossible to reach such a limit and, therefore, the measured values of  $k$  may exhibit some deviations from  $S$  or  $S'$ . Thus, such deviations are probably due to transient effects that should diminish as very low frequencies are accessed. Let us also mention that, Bergman [16] suggested that,  $S = s$ , but his suggestion is not currently supported by the estimates of the two exponents listed in Table 1.

### Physical Gels: Comparison of the Data with the Percolation Model

In physical gels both inter- and intramolecular bondings are non-covalent. The presence of non-covalent bonding means that their numbers and positions fluctuate with time, as well as temperature, as such bonds are reversible. Moreover, the nature of the (physical) crosslinks is not completely understood. In many cases they involve hydrophobic, hydrogen bonding, and electrostatic interactions, the combination of which make gaining a better understanding of the properties of physical gels a very complex problem. This is particularly true for biopolymers.

Two well-known examples of physical gels are gelation of silica particles in NaCl solutions and in pure water [36], and silica aerogels [68]. As discussed above, the attachment of the particles in such gels is by relatively weak association. The BB forces are important in such gels since touching particles that form long chains, when deformed, roll on top of one another and this motion and the displacement of the centers of any 3 mutually-touching particles create forces that are equivalent to the BB forces.

Despite the many complications, percolation theory seems to be capable of providing rational explanations for the scaling behavior of at least some of the experimental data for the elastic moduli of physical gels near the GP. For example, measurements [36,68] of the elastic moduli of silica gels and aerogels yielded an estimate of the elasticity exponent,  $z \simeq 3.8$ , which is in excellent agreement with the critical exponent  $T$  for the 3D BB model (see Table 1), as well as with Eq. (63).

More recent measurements by Devreux et al. [27] indicate a crossover between the prediction of the BB model and another regime with a much smaller value of  $z$ . They measured the complex modulus  $G^*$  of silica gels formed by hydrolysis-condensation of a silicon alkoxide. For a restricted region near the GP they reported that,  $z \simeq 2.0 \pm 0.1$ , close to the exponent  $T$  of the CF percolation, whereas beyond this region they found  $z \simeq 3.6 \pm 0.1$ , which is close to the elasticity exponent for the 3D BB model. Devreux et al. interpreted their data for the region near the GP in terms of an analogy between elastic percolation networks and random resistor networks.

### Future Directions

Percolation models explain qualitatively, and in many cases quantitatively, the structure, rheology, and mechanical behavior of branched polymers and gels. In particular, such models provide a rational explanation for the power-law behavior of many properties of such materials near the

percolation threshold, which mean-field theories and effective-medium approximations fail to predict.

Despite their success, there are still many issues to be addressed. There is still doubt as to whether percolation models can explain the behavior of many types of nearly critical gels and branched polymers, particularly biopolymers (see, for example, [56]). Several sets of puzzling data on the elastic moduli of critical and nearly critical gels remain to be explained. If there is no unique universality class for the elastic moduli and the viscosity, but there is, instead, a few of them, the crossovers between the various universality classes remain to be clarified. Our understanding of various properties of physical gels is not as extensive as what we now know about chemical gels. These and other issues promise exciting research problems for the foreseeable future.

### Bibliography

#### Primary Literature

1. Adam M, Delsanti M, Durand D (1985) Mechanical measurements in a reaction bath during the polycondensation reaction near the gelation threshold. *Macromolecules* 18:2285
2. Adam M, Delsanti M, Durand D, Hild G, Munch JP (1981) Mechanical properties near gelation threshold, comparison with classical and 3d percolation theories. *Pure Appl Chem* 53:1489
3. Adam M, Delsanti M, Munch JP, Durand D (1987) Size and mass determination of clusters obtained by polycondensation near the gelation threshold. *J Phys* 48:1809
4. Adam M, Delsanti M, Okasha R, Hild G (1979) Viscosity study in the reaction bath of the radical copolymerisation of styrene divinylbenzene. *J Phys Lett* 40:539
5. Adam M, Lairez D, Karpasas M, Gottlieb M (1997) Static and dynamic properties of cross-linked poly(dimethylsiloxane) pregel clusters. *Macromolecules* 30:5920
6. Adolf D, Martin JE, Wilcoxon JP (1990) Evolution of structure and viscoelasticity in an epoxy near the sol-gel transition. *Macromolecules* 23:527
7. Alexander S (1984) Is the elastic energy of amorphous materials rotationally invariant? *J Phys* 45:1939
8. Allain C, Limat L, Salomé L (1991) Description of the mechanical properties of gelling polymer solutions far from gelation threshold: generalized effective-medium calculations of the superconductor-conductor site percolation problem. *Phys Rev A* 43:5412
9. Allain C, Salomé L (1987) Hydrolysed polyacrylamide/Cr<sup>3</sup> gelation: Critical behaviour of the rheological properties at the sol-gel transition. *Polym Commun* 28:109
10. Allain C, Salomé L (1987) Sol-gel transition of hydrolyzed polyacrylamide + chromium III: Rheological behavior versus cross-link concentration. *Macromolecules* 20:2957
11. Allain C, Salomé L (1990) Gelation of semidilute polymer solutions by ion complexation: Critical behavior of the rheological properties versus cross-link concentration. *Macromolecules* 23:981
12. Arbabi S, Sahimi M (1990) Critical properties of viscoelasticity of gels and elastic percolation networks. *Phys Rev Lett* 65:725

13. Arbabi S, Sahimi M (1990) On three-dimensional elastic percolation networks with bond-bending forces. *J Phys A* 23:2211
14. Arbabi S, Sahimi M (1993) Mechanics of disordered solids. I Percolation on elastic networks with central forces. *Phys Rev B* 47:695
15. Axelos MAV, Kolb M (1990) Crosslinked biopolymers: Experimental evidence for scalar percolation theory. *Phys Rev Lett* 64:1457
16. Bergman DJ (2003) Exact relation between critical exponents for elastic stiffness and electrical conductivity of percolating networks. *Physica B* 338:240
17. Born M, Huang K (1954) *Dynamical Theory of Crystal Lattices*. Clarendon Press, Oxford
18. Bouchaud E, Delsanti M, Adam M, Daoud M, Durand D (1986) Gelation and percolation: swelling effect. *J Phys Lett* 47:539
19. Brinker CJ, Scherer GW (1990) *The Physics and Chemistry of Sol-Gel Processing* Academic, San Diego
20. Candau SJ, Ankrum M, Munch JP, Rempp P, Hild G, Osaka R (1985) *Physical Optics of Dynamical Phenomena in Macromolecular Systems*. Berlin, De Gruyter, p 145
21. Carmesin I, Kremer K (1988) The bond fluctuation method: A new effective algorithm for the dynamics of polymers in all spatial dimensions. *Macromolecules* 21:2819
22. Colby RH, Gilmor JR, Rubinstein M (1993) Dynamics of near-critical polymer gels. *Phys Rev E* 48:3712
23. Daoud M (1988) Distribution of relaxation times near the gelation threshold. *J Phys A* 21, L973
24. Daoud M (2000) Viscoelasticity near the sol-gel transition. *Macromolecules* 33:3019
25. Daoud M, Coniglio A (1981) Singular behaviour of the free energy in the sol-gel transition. *J Phys A* 14:301
26. Daoud M, Family F, Jannik G (1984) Dilution and polydispersity in branched polymers. *J Phys Lett* 45:199
70. de Gennes (1977) PG Critical behaviour for vulcanization processes. *J Phys Lett* 38:L355
71. de Gennes PG (1972) Exponents for excluded volume problem as derived by the Wilson method. *Phys Lett A* 38:339
72. de Gennes PG (1976) On a relation between percolation theory and the elasticity of gels. *J Phys Lett* 37:L1
73. de Gennes PG (1979) Incoherent scattering near a sol gel transition. *J Phys Lett* 40:197
74. del Gado E, de Arcangelis LE, Coniglio A (1999) Elastic properties at the sol-gel transition. *Europhys Lett* 46:288
27. Devreux F, Boilot JP, Chaput F, Malier L, Axelos MAV (1993) Crossover from scalar to vectorial percolation in silica gelation. *Phys Rev E* 47:2689
28. Djabourov M, Leblond J, Papon P (1988) Gelation of aqueous gelatin solutions. II Rheology of the sol-gel transition. *J Phys Fr* 49:333
29. Durand D, Delsanti M, Adam M, Luck JM (1987) Frequency dependence of viscoelastic properties of branched polymers near gelation threshold. *Europhys Lett* 3:297
30. Fadda GC, Lairez D, Pelta J (2001) Critical behavior of gelation probed by the dynamics of latex spheres. *Phys Rev E* 63:061405
31. Family F, Coniglio A (1980) Crossover from percolation to random animals and compact clusters. *J Phys A* 13:L403
32. Farago O, Kantor Y (2000) Entropic elasticity of two-dimensional self-avoiding percolation systems. *Phys Rev Lett* 85:2533
33. Feng S, Thorpe MF, Garboczi E (1985) Effective-medium theory of percolation on central-force elastic networks. *Phys Rev B* 31:276
34. Flory PJ (1941) Molecular size distribution in three dimensional polymers. *Gelation I. J Am Chem Soc* 63:3083
35. Gauthier MB, Guyon E (1980) Critical elasticity of polyacrylamide above its gel point. *J Phys Lett* 41:503
36. Gauthier MB, Guyon E, Roux S, Gits S, Lefaucheux F (1987) Critical viscoelastic study of the gelation of silica particles. *J Phys* 48:869
37. Gordon M, Torkington JA (1981) Scrutiny of the critical exponent paradigm, as exemplified by gelation. *Pure Appl Chem* 53:1461
38. Hodgson DF, Amis EJ (1990) Dynamic viscoelastic characterization of sol-gel reactions. *Macromolecules* 23:2512
39. Holly EE, Venkataraman SK, Chambon F, Winter HH (1988) Fourier transform mechanical spectroscopy of viscoelastic materials with transient structures. *J Non-Newtonian Fluid Mech* 27:17
40. Kantor Y, Webman (1984) I Elastic properties of random percolating systems. *Phys Rev Lett* 52:1891
41. Kirkpatrick S (1973) Percolation and conduction. *Rev Mod Phys* 45:574
42. Kremer K (1998) Numerical studies of polymer networks and gels. *Philos Mag B* 77:569
43. Lapp A, Leibler L, Schosseler F, Strazielle C (1989) Scaling behaviour of pregel sols obtained by end-linking of linear chains. *Macromolecules* 22:2871
44. Leibler L, Schosseler F (1985) Gelation of polymer solutions: an experimental verification of the scaling behavior of the size distribution function. *Phys Rev Lett* 55:1110
45. Lin YG, Mallin DT, Chien JCW, Winter HH (1991) Dynamical mechanical measurement of crystallization-induced gelation in thermoplastic elastomeric poly(propylene). *Macromolecules* 24:850
46. Lubensky TC, Isaacson J (1978) Field Theory for the statistics of branched polymers, gelation and vulcanization. *Phys Rev Lett* 41:829
47. Martin JE, Adolf D, Wilcoxon JP (1988) Viscoelasticity of near-critical gels. *Phys Rev Lett* 61:2620
48. Martin JE, Wilcoxon JP (1988) Critical dynamics of the sol-gel transition. *Phys Rev Lett* 61:373
49. Martins JL, Zunger A (1984) Bond lengths around isovalent impurities and in semiconductor solid solutions. *Phys Rev B* 30:6217
50. Parisi G, Sourlas N (1981) Critical behavior of branched polymers in the Lee-Yang edge singularity. *Phys Rev Lett* 46:891
51. Patton E, Wesson JA, Rubinstein M, Wilson JE, Oppenheimer LE (1989) Scaling properties of branched polyesters. *Macromolecules* 22:1946
52. Plischke M (2006) Critical behavior of entropic shear rigidity. *Phys Rev E* 73:061406
53. Plischke M (2007) Rigidity of disordered networks with bond-bending forces. *Phys Rev E* 76:021401
54. Plischke M, Joós B (1998) Entropic elasticity of diluted central force networks. *Phys Rev Lett* 80:4907
55. Plischke M, Vernon DC, Joós B, Zhou Z (1999) Entropic rigidity of randomly diluted two- and three-dimensional networks. *Phys Rev E* 60:3129
56. Ross MSB (2007) Biopolymer gelation-exponents and critical exponents. *Polym Bull* 58:119
57. Sahimi M (1986) Relation between the critical exponent of elastic percolation networks and the dynamical and geometrical exponents. *J Phys C* 19:79

58. Sahimi M (2003) *Heterogeneous Materials I*. Springer, New York
59. Sahimi M, Arbabi S (1993) Mechanics of disordered solids. II Percolation on elastic networks with bond-bending forces. *Phys Rev B* 47:703
60. Sahimi M, Goddard JD (1985) Superelastic percolation networks and the viscosity of gels. *Phys Rev B* 32:1869
61. Schmidt M, Burchard W (1981) Critical exponents in polymers: A sol-gel study of anionically prepared styrene-divinylbenzene copolymers. *Macromolecules* 14:370
62. Stauffer D (1976) Gelation in concentrated critically branched polymer solutions. Percolation scaling theory of intramolecular bond cycles. *J Chem Soc Faraday Trans II* 72:1354
63. Stockmayer WH (1943) Theory of molecular size distribution and gel formation in branched-chain polymers. *J Chem Phys* 11:45
64. Takahashi M, Yokoyama K, Masuda T (1994) Dynamic viscoelasticity and critical exponents in sol-gel transition of an end-linking polymer. *J Chem Phys* 101:798
65. Tokita M, Niki R, Hikichi K (1984) Percolation theory and elastic modulus of gel. *J Phys Soc Jpn* 53:480
66. Wang J (1989) The bond-bending model in three dimensions. *J Phys A* 22:291
67. Winter HH, Mours M (1997) Rheology of polymers near liquid-solid transitions. *Adv Polym Sci* 134:165
68. Woignier T, Phalippou J, Sempere R, Pelons J (1988) Analysis of the elastic behaviour of silica aerogel taken as a percolating system. *J Phys Fr* 49:289
69. X Xing, Muthupadhyay S, Goldbart PM (2004) Scaling of entropic shear rigidity. *Phys Rev Lett* 93:225701

## Books and Reviews

- Brinker CJ, Scherer GW (1990) *The Physics and Chemistry of Sol-Gel Processing*. Academic, San Diego
- de Gennes PG (1979) *Scaling Concepts in Polymer Physics*. Cornell University Press, Ithac
- Rubinstein M, Colby RH (2003) *Polymer Physics*. Oxford University Press, New York
- Stauffer D, Aharony A (1994) *Introduction to Percolation Theory*, 2nd revised ed. Taylor and Francis, London
- Sahimi M (1994) *Applications of Percolation Theory*. Taylor and Francis, London
- Sahimi M (2003) *Heterogeneous Materials I & II*. Springer, New York

## Percolation in Porous Media

PETER KING<sup>1</sup>, MOHSEN MASIHI<sup>2</sup>

<sup>1</sup> Imperial College London, London, UK

<sup>2</sup> Sharif University of Technology, Tehran, Iran

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Porous Media](#)

[Application of Percolation to Pore Scale](#)  
[Application of Percolation to the Field Scale](#)  
[Future Directions](#)  
[Bibliography](#)

## Glossary

**Anisotropy** There is anisotropy when the global physical property of the system is direction dependent.

**Breakthrough time** The time for convection of a single phase passive tracer between an injection well and a production well.

**Connectivity** The fraction of occupied sites belonging to the percolating clusters i. e. represents the strength of the percolating cluster.

**Continuum percolation** Percolation on continuum spaces with randomly distributed geometrical objects where there is no lattice at all.

**Capillary dominated flow** A flow regime in which the only dominant driving force is due to capillarity.

**Fracture** Any discontinuity within a rock mass which developed as a response to stress.

**Field scale** This represents large scale heterogeneities at reservoir level or the kilometer scale.

**Finite size scaling** A scaling law within percolation theory which deals with the effects of the finite boundaries.

**Invasion percolation** Another kind of percolation theory appropriate for describing the structure and amounts of two immiscible fluids at breakthrough.

**Modeling** Describing physical phenomena under nature's law in some mathematical relations, e. g. governing fluid flow, to better understand the system and to predict its behavior.

**Porous media** A medium consists of rock grains and disordered void spaces of approximately 10–100  $\mu\text{m}$  across usually occupied by oil, water and gas in a typical hydrocarbon reservoir and characterized by porosity and permeability.

**Pore scale** This represents pore throat level or the micron scale.

**Permeability** The “conductance” of the rock to fluid flow determined from Darcy's Law that the flow rate is proportion to the applied pressure gradient and inversely proportional to the fluid viscosity, the constant of proportionality is the permeability.

**Percolation threshold** A particular value of occupancy probability at which one large cluster spans the whole region.

**Simulation** Numerical model for solution to the mathematical equations which be able to predict the physical behavior of the system.



**Uncertainty** The estimated amount or percentage by which an observed or calculated value may differ from the true value.

### Definition of the Subject

Porous media are important in many areas including hydrocarbon reservoir engineering, hydrology and environmental engineering. They are also important in, for example, fuel cells, many industrial process and biological systems (lungs, bones, capillary networks and termite nests are all biological examples). Understanding the structure of porous media and the physics of fluid flow in porous media is of great interest. For example, choosing the efficient recovery techniques by reservoir engineers requires understanding of how different fluids and the porous media interact at different scales by simulating the fluid flow in the reservoir under variety of conditions. The exchange and transport of reagents in fuels cells governs their efficiency. Percolation theory which describes the connectivity of a system mathematically [22,100] has also many important applications from the spread of diseases and forest fires to the connectivity of geological entities (e. g., sandbodies or fractures) in porous media used for nuclear waste disposal or for hydrocarbon recovery [94]. Percolation concepts have been used to model fluid movement in porous media and fractured rocks at both pore scale (mm) and reservoir scale (km). At the pore scale a network of pore and throats is used to study the immiscible displacement and estimate the capillary pressure and relative permeabilities. At the field scale, the permeability map is split into either permeable (flow units e. g. good oil bearing sands) or impermeable (background e. g. shale) and assumes that the connectivity of flow units controls the flow movement. Then percolation theory is directly used to estimate static behavior (connectivity – i. e. connected fraction of good oil bearing sands) and dynamic behavior (i. e. effective permeability across the reservoir, breakthrough time between an injector and a producer or post breakthrough behavior). In particular, the percolation approach is able to estimate the uncertainty in the reservoir performance parameters which is not possible with a single detailed conventional simulation model.

### Introduction

Percolation theory is a mathematical model of the connectivity and conductivity in geometrically complex systems [100] first developed by Broadbent and Hammersley in [22]. The full description of this theory and its applications to different disciplines can be found elsewhere [13,

94,100]. It links the global geometrical and physical properties of the system to the number density of geometrical objects (representing geological entities, e. g., fractures) placed randomly in space through algebraic universal laws. By universality we mean the large scale behavior of the system is independent of the small scale details of the system, i. e. local geometries.

On the simplest example is an infinite lattice of sites which are occupied with a probability  $p$ . Clusters are formed as the neighboring sites are occupied and they are identified by using standard algorithm [48]. The clusters grow in size as the occupancy probability increases. Then, at particular value of  $p$ , called the percolation threshold  $p_c^\infty$ , one large cluster spans the whole region. There are also other small clusters which get absorbed as  $p$  further increases. For the infinite lattice there are some simple power law or scaling laws which describe the behavior of the system around the threshold  $p_c^\infty$  such as  $P(p) \propto (p - p_c^\infty)^\beta$  and  $\xi(p) \propto |p - p_c^\infty|^{-\nu}$  where  $P(p)$  is the probability that an occupied site belongs to the spanning cluster (so called connected fraction or connectivity) and  $\xi(p)$  is the correlation length (which is a measure of the “typical” size of the clusters, excluding the infinite cluster when above the threshold). Note that the correlation length  $\xi$  is related to the two point correlation function  $g(r)$ , which is the probability that two sites separated by a distance  $r$  are in the same cluster. The critical exponents  $\beta$  and  $\nu$  are independent of the kind of the lattice or even if there is a lattice or not (continuum system) they only depends on the dimensionality of space (i. e. 2D or 3D). Values for  $\beta = 5/36$  and 0.4 (in 2 and 3 dimensions respectively), and  $\nu = 4/3$  and 0.88 [100]. This is known as *universality* and is an important concept in percolation theory which enables us to study and understand the behavior of a very wide range of systems without needing to worry too much about the small scale detail. However, the percolation threshold does depend on the detail of the lattice.

This is the basic framework of percolation theory. To be useful in the context of reservoir modeling we need to address some issues. Everything so far has been defined for an infinite lattice. What happens if there is no lattice at all (i. e. continuum systems with randomly distributed geometrical objects) and when the system is finite. There is also the issue of how to modify these percolation laws for anisotropic systems, variable size and orientation distribution of objects and spatial correlation between objects.

### Porous Media

The flows of fluids in porous media are related to many important industrial and geological applications, such



as hydrocarbon recovery or ground water flow modeling [38,95]. The porous media are typically made from rock grains and disordered void spaces and are usually characterized by porosity  $\phi$  (i.e. the storage capacity of a rock, in other words the volume fraction which is void or pore space) and permeability  $k$  (i.e. the “conductance” of the rock to fluid flow determined from Darcy’s Law that the flow rate is proportion to the applied pressure gradient and inversely proportional to the fluid viscosity, the constant of proportionality is the permeability). The pore spaces are approximately 10–100  $\mu\text{m}$  across and are usually occupied by oil, water and gas (in a typical hydrocarbon reservoir). However, the reservoir itself may be several kilometers in extent. Fluid displacement in porous media depends on the scale and can be controlled by several forces including capillary forces (mostly at the pore scale), viscous forces and gravity forces. Hence, the type of displacement observed depends on the capillary number, which is the ratio of the viscous pressure drop at the pore scale to the capillary pressure, and the Bond number, which is the ratio of the hydrostatic pressure drop over a pore to the capillary pressure. It should be noticed that the full description of displacement process in porous media is very difficult due to the variety of physical phenomena. For example, the flow of two immiscible fluids depends on the wetting properties of the two fluids, their viscosity ratio, their respective densities, and the displacement rate. Typical flow rates in reservoirs are of order of a few feet (10’s of centimeters) a day. Hence, on the pore level the flow is controlled almost entirely by capillary forces between immiscible oil and water. However, over large distances viscous and buoyancy forces dominate [105]. To study the efficient recovery techniques (e.g. water flooding), it is necessary to simulate the fluid flow at the reservoir scale. Even with the today’s modern computers, the simulation cannot be achieved at the pore scale (typically there would be of the order of  $10^{21}$  pores in a reservoir). The conventional approach is to establish the simulation on a grid of roughly 100 m linear size which represents displacements occurring within millions of pores. The small scale physics is then represented by parameters in macroscopic partial differential equations to describe the transport of fluid in the field scale simulation. These parameters can be measured experimentally on representative core samples or they can be estimated from pore and throat network models (e.g. percolation based models).

The basic equation at the continuum scale is Darcy’s law, which states that the flow rate is proportional to the applied pressure gradient and inversely proportional to the

fluid viscosity

$$\mathbf{v} = -\frac{K}{\eta} \nabla P \quad (1)$$

where  $K$  is the rock permeability. Some analysis shows that it has dimensions of area and for typical reservoir rocks it is of the order of  $10^{-12} \text{ m}^2$ , which is about the cross sectional area of a typical pore throat (although more detailed calculations are required to obtain better estimates than this). The permeability can vary by orders of magnitude over very short differences reflecting the heterogeneity in pore size arising from the complex processes of geological deposition. This law is used in conjunction with the assumed incompressibility of the fluid to give an equation for the fluid pressure.

$$\nabla \cdot \mathbf{v} = \nabla \cdot K \nabla P = 0. \quad (2)$$

When there are two or more fluid phases present this basic law gets modified to,

$$\mathbf{v}_i = -\frac{K k_i(S_i)}{\eta_i} \nabla P_i \quad (3)$$

where the  $k_i$  are the relative permeabilities of each phase which are usually assumed to be functions of the fluid saturations ( $S$ ) only. Fluid saturation is the fraction of the volume of the pore space occupied by the phase. Along with the incompressibility condition for the total flux we also have a conservation law for each phase as,

$$\phi \frac{\partial S_i}{\partial t} + \mathbf{v}_i \cdot \nabla S_i = 0 \quad (4)$$

where  $\phi$  is the porosity. This set of equations is the most basic set of equations used to describe multi-phase flow in porous media. Typically the parameters (such as porosity, permeability and relative permeabilities) are determined empirically. There is really no microscopic averaging of the pore scale physics to determine these, although recent studies using percolation as one approach are an attempt to do this (as described in the next section). In reality more complex equations incorporating further physics (such as the phase behavior) are often included also.

The pore scale porous media can be simply modeled as a network of bonds i.e. pore throats and sites i.e. pore bodies to study the flow behavior. Then a series of displacement steps in each pore or throat are combined to simulate the flow movement. The idea was first used by Fatt [39] and since then, the capabilities of network and percolation based models have improved enormously [16,17,24,36,40,53,54,72,80,82,84,101,106]. Invasion percolation is a typical example for modeling capillary

dominated regime of flow in porous media [106] and will be more fully described in the next section. The models are then used to find capillary pressure and/or relative permeability [16,35,46,62,70,99]. For example, Soll & Celia [99] developed a pore scale capillary-dominated flow model by neglecting the viscous forces but considering the gravity effect (which modifies the local capillary pressures) to simulate capillary pressure-saturation relationships in a water-wet system. Moreover, Paterson et al. [81] used a percolation model with trapping to study the effects of spatial correlations in the pore size distributions on the relative permeabilities and residual saturations. They found lower residual saturations for correlated properties in comparison to uncorrelated ones.

At the field scale, heterogeneities which affect the flow behavior appear on all length scales from microns to tens of kilometers and have to be modeled to make reliable predictions of future reservoir performance. However, there exist very few direct measurements of the flow properties. Core plugs directly measure the permeability but they represent a volume of roughly  $10^{-13}$  of a typical reservoir. Well logs and well tests measure large volumes ( $10^{-4}$  and  $10^{-7}$  respectively) but the results have to be interpreted to infer flow properties. The flow itself takes place on the scale of the pores which are typically around  $10^{-21}$  of the volume of the reservoir. The consequence is a great deal of uncertainty about the spatial distribution of the heterogeneities which influence the flow. The conventional approach to this is to build detailed reservoir models (note that the largest of these has around  $10^7$  grid blocks so they fall very short of the actual level heterogeneity that we know about), upscale or coarse grain them to around  $10^4$  or  $10^5$  grid blocks and then run flow simulations. These models need to be taken from a whole range of possible models with a suitable probability attached to each to determine the uncertainty in performance. The problem with this approach is that it is computationally very expensive. Therefore, there is a great incentive to produce much simpler models which can predict the uncertainty in performance much quicker. These models must be based on the dominant physics that control the displacement process. The percolation approach based on the connectivity of flow units is one very quick method to model flow and predicts uncertainty in the reservoir performance parameters. This will be further described in Sect. “Application of Percolation to the Field Scale”.

### Application of Percolation to Pore Scale

Let's start with a simple case of displacement of a fluid by second fluid in a two-phase system i.e. the problem

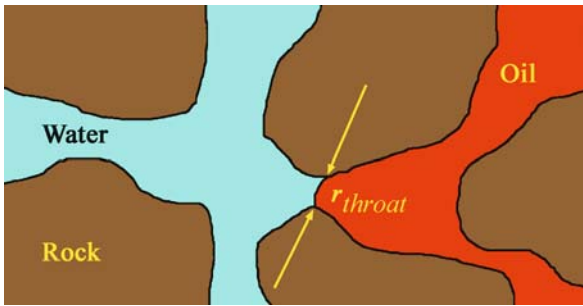
of oil/water flooding. Fluid movement can be governed by viscous, gravity and capillary forces [8]. For systems without gravity we expect different flow regimes depending on the capillary number. Viscous forces in the two fluids may be different mainly because the viscosity of the fluids is different. The high viscosity of the displaced fluid can lead to a highly unstable displacement pattern with a rapid breakthrough of the non-wetting fluid into the wetting fluid called viscous fingering [26,47]. We neglect this by considering the situation that the displacing fluid has a higher viscosity or equal viscosity than the displaced fluid. Then, for slow displacements the invasion percolation can be used to describe the structure and amounts of fluids in a two-phase displacement at breakthrough when the invading fluid is completely nonwetting [24,106] [Lenormand and Bories 1980]. The displacement in the network (or model) is based on physical principles (i.e. the heterogeneity of the capillary pressures along the interface). Consider a lattice with sites and bonds representing pores and throats respectively (with the pores as spheres and the throats as cylinders in three dimensions). The throats can be classified into allowed (those can in principle be invaded by that phase ignoring the effect of surrounding bonds), occupied (those that are occupied by that phase) and accessible (those that are allowed by the phase but also the surrounding bonds will not inhibit the fluid to try). Two processes can be considered for an immiscible displacement. An event where a wetting phase (i.e. water) is displaced by a non-wetting phase with a positive capillary pressure is called drainage. The process where the wetting phase (i.e. water) enters the porous medium and displaces the non-wetting phase is called imbibition. In practice, the capillary pressure for imbibition is lower than that for drainage.

Here we will describe how invasion percolation works for drainage. This is the simplest situation and the extension to other process can be found in the literature. First consider a simple pore with a single interface between the fluids (Fig. 1).

In order to overcome the pressure caused by the interfacial pressure to drive the non-wetting phase (oil) into the wetting phase (water) occupied pore we need to apply a pressure of

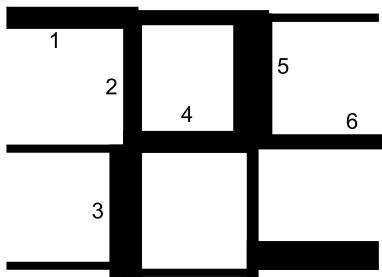
$$P_c = \frac{\gamma \cos \theta}{r_{\text{throat}}} \quad (5)$$

where  $\gamma$  is the interfacial tension,  $\theta$  is the contact angle and  $r_{\text{throat}}$  is the pore throat radius. We now imagine a network of pores linked by throats of varying radii. As we increase the pressure applied to the invading non-wetting phase it can be seen that the pores will fill from the largest first



Percolation in Porous Media, Figure 1

A schematic representation of a simple pore with a single interface between the non-wetting phase (oil) and the wetting phase (water)



Percolation in Porous Media, Figure 2

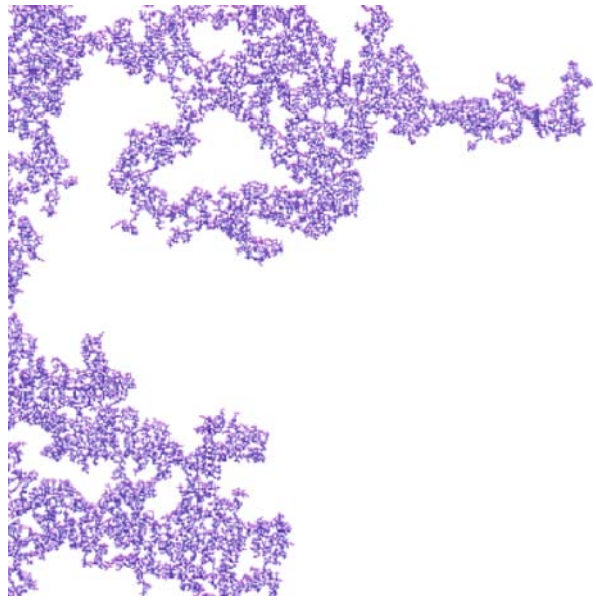
A simple illustrative network of pores linked by throats of varying radii. The numbers represent the filling order of pores

(as there the entry pressure to fill is lowest). However, the pores can be filled only if they are connected to the inlet face of in contact with the non-wetting phase. Hence, in the simple network below (Fig. 2) the pores will fill in the order 1, 2, 3 and so on where the labels are in order of the throat radius (the invading fluid comes from the left).

If this process is continued we get a not fully occupied structure (Fig. 3).

It can be seen that the displacement process is not very efficient and there are large regions unswept. This process has picked out the percolation cluster of the network. If the bonds are ordered and a threshold applied such that (starting with largest first) a fraction of bonds equal to the percolation threshold are occupied a pattern such as this will be found, except that none of the clusters not connected to the infinite cluster are found. This represents the fact that the invading fluid cannot “jump” from a current occupied site to some other interior site. Flow can only take place through sites already connected to the inlet. This is the simple model of invasion percolation as introduced by Koplik and Wilkinson [54,106].

The main quantities of interest will be the fraction of sites which become occupied by the invader, and the



Percolation in Porous Media, Figure 3

A typical percolation structure obtained from invasion percolation simulations for drainage

distribution of random numbers of those sites. In invasion percolation with no trapping the clusters exhibit fractal character with fractal dimension (see [61] for fractals)  $D = 1.89$  and  $2.52$  in 2D and 3D respectively which are similar to the results obtained from random percolation. The fraction of volume occupied by the invader is proportional to the grid size to the power  $-0.11$  and  $-0.48$  for 2D and 3D cases respectively which are again consistent with the universal values derived from random percolation results. Hence, it has the same universality class as random percolation. Invasion percolation with trapping causes the phenomenon of residual oil. Fractal dimension of invader cluster is  $1.82$  in 2D which is less than  $91/48$  of random percolation with no significant difference in 3D. The fraction of volume occupied by the invader is proportional to the grid size to power  $-0.18$  in 2D. If we pursue with the invasion process beyond the point of percolation, a second percolation threshold is reached when the defender consists of isolated clusters only and the process stops. Then the system has reached saturation of residual oil. Properties of invasion percolation are believed to be consistent with that of random percolation. However, the spanning clusters are not precisely the same. Effects of gravity were characterized through dimensionless Bond number [15,71]. Percolation theory can then be used to improve our understanding of relative permeability and capillary pressure curves in porous media (e. g. [46]).

Whilst this process is not what happens in the typical oil recovery process it is a reasonable representation of the filling of an oil reservoir when the drops of oil formed in a source rock (typically some distance and much deeper than the final reservoir rock) moves under gravity and capillary forces to displace the water that originally in the reservoir rock.

### Application of Percolation to the Field Scale

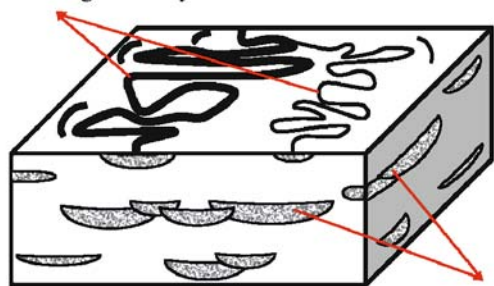
Here we describe the application of percolation theory first to low to intermediate net-to-gross conventional reservoirs and secondly to fractured reservoirs. Consider, for example, a meandering river which deposits sand over time as represented schematically in Fig. 4. The deposited sand creates a sandbody which covers the meander belt. Occasionally an event upstream, the river changes its path (called an evulsion by geologists) and deposits a new sandbody that may overlap a previous body. The process continues and forms a system of embedded sandbodies in an impermeable background. Although there may be other depositional and post-depositional events such as crevasse splays, mud drapes and shale layers which may alter this simple model, this simple model of overlapping sandbodies has been used [50,75,76] as the basic model for low to intermediate net-to-gross non fractured reservoirs.

Another example could be fracture networks. Field data obtained from large scale investigations show that the fracture network structure is close to the threshold and displays strong channeling patterns which can be explained by percolation theory [23]. It was found that in some natural fracture networks only a small percentage of fractures contributed to the permeability of the system ([95] and references therein). The large uncertainty associated with data and the lack of distinction between

faults and joints makes it debatable whether or not natural fractures are well above the percolation threshold (highly interconnected) or near the threshold (poorly connected) ([14] and references therein). As pointed out by Berkowitz [12], in some circumstances fracture networks seems to be highly connected, whereas in many other cases, where the fractures are created as a results of stress, the network is poorly connected which indicates that it is near the percolation threshold. These studies indicate that the application of percolation theory in many fracture networks is reasonable. Simple percolation models would assume that the fractures are randomly oriented and independently located in space, however, in reality fractures show different orientation distribution [1,9,95,108] result from successive tectonic events and several length distributions such as power law [45,77,97,102], log normal (e.g. [77,93]) or exponential (e.g. [86,93]). There also exist spatial correlation between fractures ([18], and references therein) and cross correlation between fracture parameters such as correlation between aperture and the fracture size (e.g. [79,103]) or cross correlation between the position of a fracture and its length (e.g. [21,29]). We shall first discuss the very simple model of fractures [4,5,11,19,20,25,65,89] made by constant length randomly oriented and/or orthogonal fractures (Fig. 5).

In these two examples, the objects (i. e. sandbodies and fractures) are not restricted to points on a fixed lattice

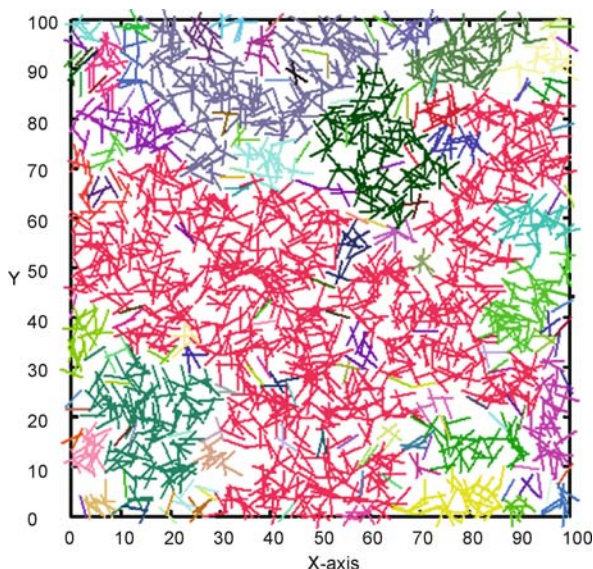
Meandering River System



Deposited Sandbodies

Percolation in Porous Media, Figure 4

A meandering river system deposits sandbodies over the ages [75]



Percolation in Porous Media, Figure 5

A randomly oriented fracture network ( $l = 5$ ) at the threshold with 2066 fractures where the spanning cluster shown in red consists of 897 fractures



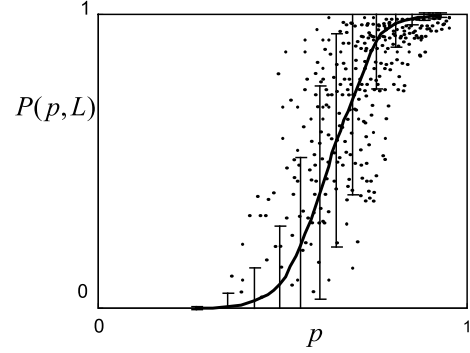
so as there is no maximal concentration (notice that the upper bound of the occupancy probability in simple lattice percolation is one). In the case of fracture systems, for instance, there is theoretically no end to the degree of fracturing [95] and fractures can be of any shape with variable length, direction and number of interconnected bonds [11]. These lead us to use percolation on continuum spaces for these cases instead of using percolation on lattices.

### Continuum Percolation

This is very straightforward because of the universality principle. We can place geometrical objects (e.g. rectangles representing sandbodies or line segments representing fractures in 2D space) randomly and independently (so called a *Poisson process*) in space. In place of the occupancy probability  $p$  we have the volume fraction of objects (or the probability that a point chosen at random lies within one of the objects) with the same notation,  $p$ . We get the same threshold phenomenon of a single cluster growing and dominating the system. The percolation threshold depends only on the shape of the objects, but for circles it is  $0.678 \pm 0.0024$  and for squares it is  $0.668 \pm 0.0026$  (similarly in 3D for spheres it is  $0.288 \pm 0.0016$  and for cubes  $0.276 \pm 0.0013$ ) so the difference is not very large and numerical experiments indicate that for reasonable convex (i.e. not very spiky) objects the threshold is around the same value. This is known as *continuum percolation*. Examples of applying percolation theory to uncorrelated (or even correlated) continuum systems that check the universality and determination of the percolation threshold of different models can be found elsewhere (e.g. [3,11,27,28,41,42,50,57,59,60,107]). Extensive studies have shown that fracture systems, for example, belong to the general class of continuum percolation systems (e.g. [1,5,11,55]). From the principle of universality, the critical exponents are then fixed but the percolation threshold depends on the network topology. Previous estimations of critical exponents were successfully close to those from lattice percolation (e.g. [1,5,19,20,25,90]). Hence, from the principle of universality, we can use the same scaling laws with the same numerical values of the critical exponents as in lattice percolation. This is a remarkable result that we can now use.

### Finite Size Scaling

The problem of how to deal with finite size systems is known as *finite size scaling*. The simple scaling laws described in Sect. “[Introduction to Percolation Theory](#)” only apply to infinite-size systems. In a finite system because



**Percolation in Porous Media, Figure 6**

A typical scatter with the curve and the lines represent respectively the mean  $P(p, L)$  and the standard deviation  $\Delta(p, L)$  of connected fraction determined over all realizations at the same occupancy probability of a finite size system

of a sample-size uncertainty there may be a connection at an occupancy very much less than the threshold value but still no connection at very high  $p$  values (greater than the threshold). This makes the definition of the percolation threshold of a finite system unclear, as a spanning cluster may appear in one realization and not in another at the same  $p$ . Therefore, for a finite system, an apparent threshold  $\tilde{p}_c(L)$  which depends on the system size can be used. There have been several definitions of the apparent threshold in the literature [1,11,50,100] such as the occupancy probability at which half of the realizations percolate. The percolation probability (i.e. connectivity)  $P(p, L)$  can be defined as the fraction of occupied sites belonging to the spanning clusters. This is the finite size analogue of  $P(p)$ . If we plot  $P(p, L)$  as a function of  $p$  over a large number of realizations for a particular system size (Fig. 6), we get a scatter of points from which we can determine the mean connected fraction  $P(p, L)$  (the same notation as before) and the standard deviation  $\Delta(p, L)$  (the fluctuations about this mean value).

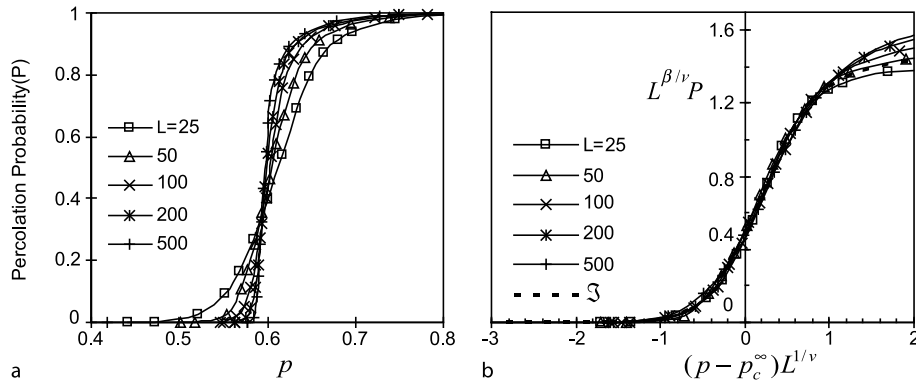
The effect of finite boundaries is to *smear out* the percolation transition (there is not a sharp transition in the connectivity as was exist in infinite systems). Plotting the mean connectivity  $P(p, L)$  and the standard deviation of connectivity  $\Delta(p, L)$  results obtained from different system sizes as a function of  $p$  gives different curves (Fig. 7a) which can be related to each other through the finite-size scaling law [100]:

$$P(p, L) = L^{-\beta/\nu} \mathfrak{F}[(p - p_c^\infty)L^{1/\nu}] \quad (6)$$

$$\Delta(p, L) = L^{-\beta/\nu} \mathfrak{R}[(p - p_c^\infty)L^{1/\nu}] \quad (7)$$

where  $\mathfrak{F}$  and  $\mathfrak{R}$  are two scaling functions for the mean and standard deviation of connectivity, respectively.





**Percolation in Porous Media, Figure 7**

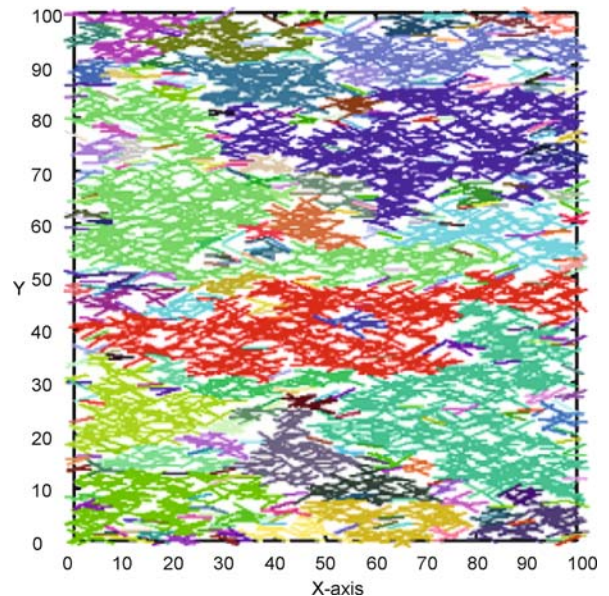
Plot of the mean connected fraction  $P(p, L)$  showing: a the effects of finite boundaries on the percolation transition and b the data collapse using finite size scaling transformations

This means that, for example, if we plot  $L^{\beta/\nu} P$  against  $(p - p_c^\infty) L^{1/\nu}$  all the mean connectivity curves found previously should lie on top of each other to form a single universal curve  $\mathfrak{F}$  (Fig. 7b).

It has been shown that all of the curves lie nearly on top of each other except as  $p$  approaches unity [65,75,76]. This corresponds to the region where simple scaling breaks down. This region may be small in some cases and could be treated by using effective medium theory [43,50,52,87,98]. It should be noticed that the scaling laws in Eqs. (6) and (7) are universal, but the scaling curves  $\mathfrak{F}$  and  $\mathfrak{R}$  depend on the model. These are very useful results, because once we get the two scaling curves from numerical simulations for a specific model, we can quickly predict the mean connectivity and its associated uncertainty for any other system sizes without performing any explicit realizations. Examples of the scaling master curves for the mean connectivity and the standard deviation of connectivity for fracture model and sandbodies model were given in Masihi et al. [65,66], and Nurafza et al. [75,76], respectively.

### Anisotropy

The other problem that we have to deal with is due to anisotropy. By isotropy we mean that the horizontal connectivity is the same as the vertical connectivity on average if not for individual realizations. However, for many realistic systems, the objects or their orientation are rarely isotropic. For example in fractured rocks fracture sets with particular orientations are typically formed as a result of tectonic history [1,9,43,95,108]. This leads to the creation of an *easy* direction for connected paths which is in the short direction and a *difficult* direction which is along the long axis.



**Percolation in Porous Media, Figure 8**

An anisotropic fracture network ( $l = 5$ ) at the threshold with random orientation in the range  $\theta \in (-30^\circ, 30^\circ)$  around the horizontal showing the *easy* direction for connection along x-axis with 3802 fractures where the percolating cluster consists of 432 fractures. Connectivity along the y-axis, which is the *difficult* direction for connection, needs much more fractures to be placed in the regions

The question is how to apply finite size scaling to anisotropic systems. This requires understanding the anisotropic behavior in percolation. A survey of the literature shows few studies on the subject among which are Monetti and Albano [73] who performed Monte-Carlo simulations in an elongated geometry to obtain the dependency of the horizontal and vertical finite size perco-

lation threshold to the aspect ratio of the lattice; Marrink and Knackstedt [63] who assumed that an elongated lattice can be treated as a series of linked isotropic lattices; Hovi and Aharonyt [49] who used renormalization group theory and duality arguments; Langlands et al. [56] who found numerically the dependency of the crossing probability on the aspect ratio of rectangular systems.

Recently, Masihi et al. [68], have shown that it is possible to account for moderate anisotropy in finite size scaling within percolation by first using the apparent threshold  $\tilde{p}_c$  in the principal coordinate directions of the anisotropy as the value of  $p$  when 50% of realizations connect in that direction instead of the infinite percolation threshold and then rescaling with the effective length  $L_x$  as,

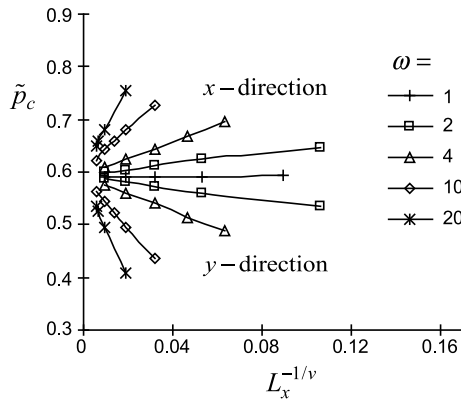
$$P(p, L_x, \omega) = L_x^{-\beta/\nu} \mathfrak{F}[(p - \tilde{p}_c)L_x^{1/\nu}] \quad (8)$$

$$\Delta(p, L_x, \omega) = \omega^{1/2} L_x^{-\beta/\nu} \mathfrak{R}[(p - \tilde{p}_c)L_x^{1/\nu}] \quad (9)$$

where aspect ratio  $\omega = L_x/L_y$  represents the anisotropy. The apparent threshold in each direction is given by,

$$\tilde{p}_c^i = p_c^\infty + \Lambda_i(\omega)L_x^{-1/\nu} \quad (10)$$

which has a symmetry property (see Fig. 9) where the constant of proportionality is  $\Lambda(\omega) = c(\omega^{1/\nu} - 1)$  with  $c \approx 0.92, 0.58$  and  $0.41$  for respectively elongated lattice [68], anisotropic fracture model [69] and anisotropic sandbody model [75,76]. This means that we can use the same isotropic universal curves ( $\mathfrak{F}$  and  $\mathfrak{R}$ ) for predicting connectivity of anisotropic cases.



Percolation in Porous Media, Figure 9

Plot of apparent threshold in both the  $x$  and the  $y$  directions of anisotropic lattices as a function of  $L_x^{-1/\nu}$ , showing that the shift in the apparent thresholds is symmetrically placed about the isotropic case ( $\omega = 1$ )

## Size Distribution

Another key parameter affecting the connectivity is the size distribution. In reality the sandbodies, for example, may have different sizes based on the sedimentological environment in which the sands were deposited. Also fractures usually have a length distribution depending on the degree of rock deformation (e.g. [88]) from negative-exponential and log-normal to power-law distributions (e.g. [18,45,78,97,102]).

The analysis of the connectivity based on finite size scaling that we have discussed so far assumes that objects (i.e. fractures or sandbodies) all have the same lengths. However the distribution of sizes introduces a new complication. The idea behind the finite size scaling is that the percolation behavior is controlled by two dimensionless lengths, the system size,  $L$ , and the correlation length,  $\xi$  (these are made dimensionless by scaling with the linear dimension of the geometrical object). If there is a distribution of lengths then this is changed. One might expect that the connectivity behavior of such a system with a distribution of lengths is identical to the connectivity behavior of a system with constant-length objects whose object length (called the effective length,  $l_{\text{eff}}$ ) can be defined in an appropriate way. In the case of fracture model, for example, using the concept of effective length introduced by Robinson [89,90] and Balberg et al. [6], the two previously determined universal connectivity curves for the constant length fractures ( $\mathfrak{F}$  and  $\mathfrak{R}$ ) can be applied to fracture systems with a distribution of fracture lengths. This representative length can be based on either the first moment  $\langle l \rangle$  or the second moment  $\langle l^2 \rangle$  of the fracture length distribution [6,11,19,66,74,89,91]. From numerical studies Masihi et al. [66] found that the second moment gave a better fit to the data. In the case of very wide distribution of lengths (i.e. power law) the scaling exponents may be different from the standard values as there is a possibility for a single large fracture to connect both sides of the system and dominates the connectivity. In other words, in this case there is a break down in universality for the critical exponents. The dependency of the scaling exponents of percolation theory on the exponent of power law length distribution has been investigated numerically [19,66].

## Orientation Distribution

In reality the sandbodies, for example, are not all aligned in one direction and are affected by the depositional process. Note that the orientational disorder of the bodies which is more relevant in three dimensions can greatly enhance the connectivity of the system, particularly for the systems with thin long bodies. In percolation terminology this is

a reduction in threshold which is not due to the finite size effects but a real shift in the infinite threshold. Very thin and long bodies, for example, if they are aligned they will not intersect so  $p_c \approx 1$  otherwise if they have an angular dispersion they will intersect at any fractional concentration and so  $p_c = 0$ . There will also be another shift in apparent percolation due to finite size effects. Nurafza et al. [75,76] studied these effects numerically and showed that the only effect of the orientational disorder is to make the bodies appear a bit larger than they are and a bit less elongated. They defined a new aspect ratio and used the reduced percolation thresholds to account for the effects of orientation of sandbodies within the finite size scaling laws which makes the previously determined universal curves applicable for orientated sandbodies.

### Spatial Correlation

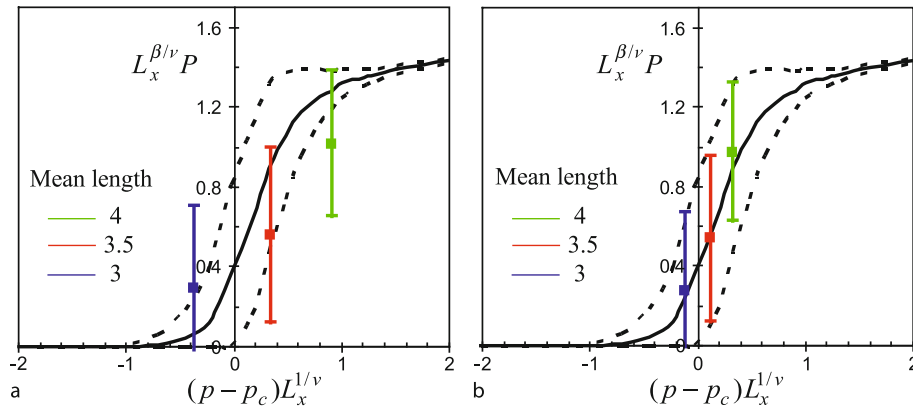
The models discussed so far assumed the objects are distributed randomly and independently in the region (Poisson statistics). This is somehow in contradiction with the known existence of, for example, fracture sets or observed spatial correlation between natural fractures over scales (e.g. [18]). A survey of the literature shows that there has been little investigation of the percolation properties of systems with short- and long range correlations. Harter [44] has shown that the percolation threshold in Markov chain random field with short range correlation decreases as the correlation scale increases but the critical exponents are unaffected (they belong to the same universality class as uncorrelated percolation). For systems with long range correlations, on the other hand, it has been shown that the percolation behavior is drastically changed as even the critical exponents may become different [85,96].

In the study of fracture networks an investigation of the spatial correlation of fractures has concentrated on the long range fracture density correlations modeled by fractal geometry [14,31,33] [Watanabe and Takahashi, 1995] showed that the correlation pattern is likely to affect the connectivity behavior. When dealing with fracture correlation it is not straightforward to find the right percolation parameter  $p$  which is able to measure the connectivity of the fracture network. As pointed out by Darcel et al. [31] neither the mean fracture density  $\rho = Nl^2/L^2$ , proposed by Bour and Davy [19], nor the mean fractal fracture density  $\rho_D = Nl^D/L^D$ , suggested by Berkowitz et al. [14] are able to represent the connectivity state of constant-length fractal fracture networks. Darcel et al. [31] also emphasized that the transition width at the threshold for a large system with fractal correlation remains fixed (it

does not vanish) which is in contrast to the second order phase transitions of percolation theory. Recently, Masihi and King [64] have presented a model of fractures which used a simulated annealing algorithm (with an objective function defined by the spatial correlation in the displacement of fractures) to generate realizations of correlated fracture networks and then used them in the percolation approach to investigate the effects of fracture spatial correlation. They have found that the scaling exponents of the connectivity are different from the conventional, uncorrelated values (see Masihi et al. [64,66] for more details).

With this background we now describe the percolation framework to model field scale reservoirs. We start with conventional reservoirs. Hydrocarbon reservoirs have a complicated geometry due to the complex sedimentary processes deposited them over the years. They consist of good sandstone (i.e. high permeability and porosity) containing oil within their pores, and poorer siltstones, mudstones and shales (i.e. low permeability). The main flow of oil during recovery is through these good sands and flow through poorer rock being too slow to be of economic consequence. Hence, connectivity of these sandbodies (also called flow units) across the reservoir or in between an injection and a production well is crucial. Note that the total sands gives the total oil in place and the fraction of connected sands between two wells shows the expected recoverable oil between the two wells. A part of the connected sand fraction is dead end and cannot contribute to the flow. The flowing part (backbone) of connected sands controls the flow through sands and so affects the effective permeability, sweep efficiency and breakthrough and post break through time behavior. Many important decisions for a given field can be made based on the knowledge of the connectivity and conductivity of these flow units. For example, at the exploration and appraisal stage decisions about initial well spacing and location can be made based on oil volumes connected by the wells and recovery factors; during plateau phase the decision about the end of plateau and the rate of increases of water or gas ratios depends on the knowledge of the geometry of the backbone or during the decline phase decisions about targeting infill wells to extend field life will be based on volumes of unswept oil or uncontacted oil.

The conventional approach to address these entails building detailed reservoir models which is very expensive in terms of human and CPU times. It has long been understood that flow in heterogeneous porous media is largely controlled by continuity of permeability contrasts, either flow barriers (e.g. shales or high permeability streaks) or faults. Although there are other influences these are the



**Percolation in Porous Media, Figure 10**

**A comparison of connectivity scaling of three correlated fracture networks with the average fracture length 4, 3.5 and 3 with the universal connectivity curves (solid and dotted curves are  $P$  and  $P \pm \Delta$ ) using: a standard values for the exponent ( $1/\nu = 0.75$  and  $\beta/\nu = 0.105$ ) and **b** modified values  $1/\nu = 0.4$  and  $\beta/\nu = 0.09$**

predominant features affecting flow. With this in mind we look to model reservoir flow which concentrates on the connectivity of permeability contrasts.

Imagine a typical reservoir model constructed with an object based technique [34]. That is geometrical objects (representing geological entities, e.g. shales, fractures, sand bodies etc.) are placed randomly in space. For example, the reservoir may have been deposited by meandering river belts in which case the good sand occurs as packages in a low permeability background. Then the connectivity and conductivity can be estimated directly by percolation theory. Note that the net to gross ratio is the volume fraction of the good sand and is, therefore, identical to the occupancy probability  $p$ . Suppose we have a reservoir of size  $L$  and a pair of wells separated by a Euclidean distance  $r$ . We can ask questions about the probability that the two wells are connected, or in percolation terminology, in the same cluster. This is just the two point correlation function defined previously. Suppose we want to know what fraction of the sand in contact with the wells is connected to both wells. This is just the connectivity function  $P$  defined earlier. We can use finite size scaling to estimate this fraction. Also we can use related scaling laws to estimate the uncertainty. Note that these are algebraic laws with no spare parameters. The percolation threshold is defined by the shape of the objects, but it is largely unimportant whether we model the sand units as rectangles or ellipsoids or other shapes (provided they are not too exotic). The scaling laws and exponents are determined from lattice models (and this has been done very extensively in the literature) and can be straightforwardly applied.

The percolation framework can be used to find the probability distribution for the breakthrough time for con-

vection of a single phase passive tracer between an injector and a producer. Comparison of preliminary results [2,37,51,58] showed that the agreement between prediction from scaling law and the numerical simulation which is good enough for engineering purposes given the fact that the prediction from the scaling law took a fraction of a second of CPU times compared with the hours required for the conventional detailed simulation. Having estimated the breakthrough time, Roslien et al. [92] showed that log-log plot of the mean and variance in production results condition to any breakthrough time  $t_{br}$  against  $(t - t_{br})$  will lie on top of each other to produce two universal curves. Now with these master curves, one can make a rapid estimate of the mean and variance in future production (e.g. the time taken for production to fall by 50% or water cut to increase to 50%) good enough for engineering purposes.

In fractured reservoirs with low matrix permeability the flow behavior depends strongly on the spatial distribution of the fractures. In this case, the fractures are the flow units which need to be distributed randomly in the impermeable background (i.e. matrix). Then again the connectivity and conductivity can be estimated directly by percolation theory. The occupancy  $p$  can be interpreted as the density of fractures e.g. number density or the critical fracture length necessary to ensure percolation for a given number of fractures in a domain [83]. Equivalent terms to this can be volumetric base (i.e., average number of fractures in the region), topological base [1] (i.e., average number of connections with surrounding fractures) or a combination [67] based on average excluded area i.e., the area around a fracture in which the center of another fracture must lie in order for them to overlap over the distri-



bution of the fracture orientation and [6] length. Extensive studies have shown that constant-length fracture models belong to the general universality class of continuum percolation systems [5,11,19,20,25,90]. Recently, Belayneh et al. [10] have applied scaling laws from percolation theory to predict the connectivity of mineralized fractures with length distribution exposed on the southern margin of the Bristol Channel Basin.

### Future Directions

There is obviously a need for further development to turn many concepts involved in this article into practical application. One area is to look for connectivity between two wells (represented by points or lines in 2D and 3D, respectively) similar to the previous works on the structure of the cluster connecting two given sites or lines of a 2D and 3D lattices [7,32]. Clearly the connectivity between two points will be lower than that between two sides. In determining the connectivity between two points the correlation function,  $g(r)$  defined previously has a major role.

Percolation can do more than predict static connectivity. There are scaling laws for the effective permeability  $K_{\text{eff}}(p - p_c^\infty)^\mu$  for infinite size systems where the conductivity exponent  $\mu$  is about 1.3 and 2 in two and three dimensions respectively [100]. It is clear that only a subset of the percolating cluster (called the backbone) can be swept whereas the dead ends are stagnant. It would be expected to see a similar finite size scaling and anisotropy effects to those for the connectivity. However, the detailed development for this is computationally very demanding as it requires solving the flow equations.

### Bibliography

- Adler PM, Thovert JF (1999) Fractures and fracture networks. Kluwer, London
- Andrade JS, Buldyrev SV, Dokholyan NV, Havlin S, Lee Y, King PR, Paul G, Stanley HE (2000) Flow between two sites in percolation systems. *Phys Rev E* 62:8270–8281
- Baker R, Paul G, Sreenivasan S, Stanley HE (2002) Continuum percolation threshold for interpenetrating squares and cubes. *Phys Rev E* 66:046136
- Balberg I (1985) Universal percolation threshold limits in the continuum. *Phys Rev B* 31(6):4053–4055
- Balberg I, Binenbaum N, Anderson CH (1983) Critical behaviour of the two dimensional sticks system. *Phys Rev Lett* 51(18):1605–1608
- Balberg I, Anderson CH, Alexander S, Wagner N (1984) Excluded volume and its relation to the onset of percolation. *Phys Rev B* 30(7):3933–3943
- Barthelemy M, Buldyrev SV, Havlin S, Stanley HE (1999) Scaling for the critical percolation backbone. *Phys Rev E* 60(2):R1123–R1125
- Bear J (1972) Dynamics of fluids in porous media. American Elsevier Publishing Company, New York
- Bear J, Tsang CF, de Marsily G (1993) Flow and contaminant transport in fractured rock. Academic Press, San Diego, pp 169–231
- Belayneh M, Masihi M, King PR, Matthäi SK (2006) Prediction of fracture uncertainty using percolation approach: model test with field data. *J Geophys Eng* 3:219–229
- Berkowitz B (1995) Analysis of fracture network connectivity using percolation theory. *Math Geol* 27(4):467–483
- Berkowitz B (2002) Characterizing flow and transport in fractured geological media: a review. *Adv Water Resour* 25:861–884
- Berkowitz B, Balberg I (1993) Percolation theory and its application to ground water hydrology. *Water Resour Res* 29(4):775–794
- Berkowitz B, Bour O, Davy P, Odling N (2000) Scaling of fracture connectivity in geological formations. *Geophys Res Lett* 27(14):2061–2064
- Birovljev A, Furuberg L, Feder J, Jssang T, Mly KJ, Aharony A (1991) Gravity invasion percolation in two dimensions: Experiment and simulation. *Phys Rev Lett* 67:584–587
- Blunt MJ, King PR (1990) Macroscopic parameters from simulations of pore scale flow. *Phys Rev A* 42(12):4780–4789
- Blunt MJ, King M, Scher H (1992) Simulation and theory of two phase flow in porous media. *Phys Rev A* 46(12):7680–7699
- Bonnet E, Bour O, Odling NE, Davy P, Main I, Cowie P, Berkowitz B (2001) Scaling of fracture systems in geological media. *Rev Geophys* 39(3):347–383
- Bour O, Davy P (1997) Connectivity of random fault networks following a power law fault length distribution. *Water Resour Res* 33(7):1567–1583
- Bour O, Davy P (1998) On the connectivity of three dimensional fault networks. *Water Resour Res* 34(10):2611–2622
- Bour O, Davy P (1999) Clustering and size distributions of fault pattern: theory and measurements. *Geophys Res Lett* Press, 26:2001–2004
- Broadbent I, Hammersley JM (1957) Percolation processes 1. Crystals and mazes. *Proc Camb Philos Soc* 53:629–641
- Cacas MC, Ledoux E, de Marsily G, Tillie B, Barbreaux A, Durand E, Feuga B, Peaudecerf P (1990) Modeling fracture flow with stochastic discrete fracture network: calibration and validation, 1. the flow model. *Water Resour Res* 26(3):479–489
- Chandler R, Koplik J, Lerman K, Willemsen JF (1982) Capillary displacement and percolation in porous media. *J Fluid Mech* 119:249–267
- Charlaix E (1986) Percolation threshold of random array of discs: a numerical simulation. *J Phys A Math Gen* 19(9):L533–L536
- Chen JD, Wilkinson D (1985) Pore-scale viscous fingering in porous media. *Phys Rev Lett* 55:1892–1895
- Choi HS, Talbot J, Tarjus G, Viot P (1995) Percolation and structural properties of particle deposits. *Phys Rev E* 51(2):1353–1363
- Consiglio R, Zouain RNA, Baker DR, Paul G, Stanley HE (2004) Symmetry of the continuum percolation threshold in systems of two different size objects. *Physica A* 343:343–347
- Darcel C, Bour O, Davy P (2003) Cross-correlation between length and position in real fracture networks. *Geophys Res Lett* 30(12):52.1–52.4



30. Darcel C, Bour O, Davy P (2003) Stereological analysis of fractal fracture networks. *J Geophys Res* 108(B9):ETG13.1–ETG13.14
31. Darcel C, Bour O, Davy P, De Dreuzy JR (2003) Connectivity properties of two dimensional fracture networks with stochastic fractal correlation. *Water Resour Res* 39(10):SBH1.1–SBH1.13
32. Da Silva LR, Paul G, Havlin S, Baker DR, Stanley HE (2003) Scaling of cluster mass between two lines in 3d percolation. *Physica A* 318:307–318
33. De Dreuzy JR, Darcel C, Davy P, Bour O (2004) Influence of spatial correlation of fracture centres on the permeability of two-dimensional fracture networks following a power law length distribution. *Water Resour Res* 40:W01502.1–W01502.11
34. Deutsch CV (2002) *Geostatistical Reservoir Modeling*. Oxford University Press, New York
35. Diaz CE, Chatzis I, Dullien FAL (1987) Simulation of capillary pressure curves using bond correlated site percolation on a simple cubic network. *Transp Porous Media* 2:215–240
36. Dijkstra T, Bartelds GA, Bruining J, Hassanizadeh M (1999) Dynamic Pore-Scale network for Two-Phase Flow. In: *Proceedings on the international workshop on characterization and measurement of hydraulic properties of unsaturated porous media*, Riverside, pp 63–71
37. Dokholyan NV, Lee Y, Buldyrev SV, Havlin S, King PR, Stanley HE (1998) Scaling of the distribution of shortest paths in percolation. *J Stat Phys* 93:603–613
38. Dullien FAL (1992) *Porous media fluid transport and pore structure*, 2nd edn. Academic Press, San Diego
39. Fatt (1956) The network model of porous media III. dynamic properties of networks with tube radius distribution. *Trans Amer Inst Min, Metall, Petrol Eng* 207:164–181
40. Fenwick DH, Blunt MJ (1998) Three-dimensional modeling of three phase imbibition and drainage. *Adv Water Resour* 21(2):121–143
41. Garboczi EJ, Snyder KA, Douglas JF, Thorpe MF (1995) Geometrical percolation threshold of overlapping ellipsoids. *Phys Rev E* 52(1):819–827
42. Gawlinski ET, Stanley HE (1981) Continuum percolation in two dimensions: Monte Carlo tests of scaling and universality for non-interacting discs. *J Phys A Math Gen* 14:L291–L299
43. Harris CK (1992) Effective medium treatment of flow through anisotropic fracture system-Improved permeability estimates using a new lattice mapping. *Trans Porous Media* 9:287–295
44. Harter T (2005) Finite size scaling analysis of percolation in three dimensional correlated binary Markov chain random fields. *Phys Rev E* 72:026120.1–26120.7
45. Heffer KJ, Bervan TG (1990) Scaling relationships in natural fractures-data, theory and applications. Paper SPE 20981 presented at Europec conference, The Hague, 20–24 Oct
46. Heiba AA, Sahimi M, Scriven LE, Davis HT (1992) Percolation theory of two-phase relative permeability. *SPE Reserv Eng* 7:123–132
47. Homsy GM (1987) Viscous fingering in porous media. *Annu Rev Fluid Mech* 19:271–311
48. Hoshen J, Kopelman R (1976) Percolation and cluster distribution, I, cluster multiple labelling technique and critical concentration algorithm. *Phys Rev B* 14(8):3438–3445
49. Hovi J-P, Aharony A (1996) Scaling and universality in the spanning probability for percolation. *Phys Rev E* 53:235
50. King PR (1990) The connectivity and conductivity of overlapping sandbodies. In: Buller AT (ed) *North Sea Oil and Gas Reservoirs II*. Graham and Trotman, London, pp 353–361
51. King PR, Buldyrev SV, Dokholyan NV, Havlin S, Lee Y, Paul G (2001) Predicting oil recovery using percolation theory. *Petrol Geosci* 7:S105–S107
52. Kirkpatrick S (1973) Percolation and conduction. *Rev Mod Phys* 45:574–588
53. Knackstedt MA, Sheppard AP, Sahimi M (2001) Pore network modelling of two-phase flow in porous rock: the effect of correlated heterogeneity. *Adv Water Resour* 24(3–4):257–277
54. Koplik J, Lasseter TJ (1985) Two-phase flow in random network models of porous media. *SPE J* February:89–100
55. Koudine N, Garcia RG, Thovert JF, Adler PM (1998) Permeability of three dimensional fracture networks. *Phys Rev E* 57(4):4466–4479
56. Langlands RP, Pichet C, Pouliot P, Saint Aubin Y (1992) On the universality of crossing probabilities in two-dimensional percolation. *J Stat Phys* 67:553
57. Lee SB, Torquato S (1990) Monte Carlo study of correlated continuum percolation: Universality and percolation thresholds. *Phys Rev A* 41(10):5338–5344
58. Lee Y, Andrade JS, Buldyrev SV, Dokholyan NV, Havlin S, King PR, Paul G, Stanley HE (1999) Traveling time and traveling length in critical percolation clusters. *Phys Rev E* 60:3425–3428
59. Lenormand R, Bories S (1980) *CR Acad Sci, Paris* B291:279
60. Lin C-Y, Hu C-K (1998) Universal finite size scaling functions for percolation on three dimensional lattices. *Phys Rev E* 58(2):1521–1527
61. Lorenz CD, Ziff RM (2001) Precise determination of the critical percolation threshold for the three dimensional Swiss cheese model using a growth algorithm. *J Chem Phys* 114(8):3659–3661
62. Mandelbrot BB (1982) *The fractal geometry of nature*. W.H. Freeman, New York, pp 468
63. Mani V, Mohanty KK (1998) Pore-level network modeling of three-phase capillary pressure and relative permeability curves. *SPE J* 3:238–248
64. Marrink SJ, Knackstedt MA (1999) Percolation thresholds on elongated lattices. *J Phys A Math Gen* 32:L461–L466
65. Masihi M, King PR (2007) A correlated fracture network: modeling and percolation properties. *Water Resour Res* 43. doi: 10.1029/2006WR005331
66. Masihi M, King PR, Nurafza P (2005) Fast estimation of performance parameters in fractured reservoirs using percolation theory. Paper SPE 94186 presented at the 14th Biennial Conference, Madrid, 13–16 June
67. Masihi M, King PR, Nurafza P (2006) Connectivity prediction in fractured reservoirs with variable fracture size; analysis and validation. Paper SPE 100229 presented at the SPE Europec, Vienna, 12–15 June
68. Masihi M, King PR, Nurafza P (2006) Connectivity of fracture networks: the effects of anisotropy and spatial correlation. Paper CMWR XVI-99 presented at the Computational Methods in Water Resources conference, Copenhagen, 19–22 June
69. Masihi M, King PR, Nurafza P (2006) The Effect of Anisotropy on Finite Size Scaling in Percolation Theory. *Phys Rev E* 74:042102

69. Masihi M, King PR, Nurafza P (2007) Fast estimation of connectivity in fractured reservoirs using percolation theory. *SPE J* 12(2):167–178
70. Maximenko A, Kadet VV (2000) Determination of relative permeabilities using the network models of porous media. *J Petrol Sci Eng* 28(3):145–152
71. Méheust Y, Løvoll G, Måløy KJ, Schmittbuhl J (2002) Interface scaling in a 2d porous medium under combined viscous, gravity and capillary effects. *Phys Rev E* 66:51603–51615
72. Mohanty KK, Salter SJ (1982) Multiphase flow in porous media: II Pore-level modeling. Paper SPE 11018, Proceedings of the 57th SPE Annual Fall Technical Conference and Exhibition, New Orleans, 26–29 Sept
73. Monetti RA, Albano EV (1991) Critical behaviour of the site percolation model on the square lattice in a  $L \times M$  geometry. *Z Phys B Condensed Matter* 82:129–134
74. Mourzenko VV, Thovert JF, Adler PM (2005) Percolation of three dimensional fracture networks with power law size distribution. *Phys Rev E* 72:036103.1–036103.14
75. Nurafza P, King PR, Masihi M (2006) Facies connectivity modelling; analysis and field study. Paper SPE 100333 presented at the SPE Europec, Vienna, 12–15 June
76. Nurafza P, Masihi M, King PR (2006) Connectivity modeling of heterogeneous systems: analysis and field study. Paper CMWR XVI-189 presented at the Computational Methods in Water Resources conference, Copenhagen, Denmark, 19–22 June
77. Odling NE (1997) Scaling and connectivity of joint systems in sandstones from western Norway. *J Struct Geol* 19(10):1257–1271
78. Odling NE, Gillespie P, Bourguin B, Castaing C, Chiles J-P, Christensen NP, Fillion E, Genter A, Olsen C, Thrane L, Trice R, Aarseth E, Walsh JJ, Watterson J (1999) Variations in fracture system geometry and their implications for fluid flow in fractured hydrocarbon reservoirs. *Petrol Geosci* 5:373–384
79. Olson JE (2003) Sublinear scaling of fracture aperture versus length: an exception or the rule? *J Geophys Res* 108(B9):ETG3.1–ETG3.11
80. Øren PE, Bakke S, Arntzen OJ (1998) Extending predictive capabilities to network models. *SPE J* 3(4):324–336
81. Paterson L, Lee JY, Pinczewski WV (1997) Three-phase relative permeability in heterogeneous formations. Paper SPE 38882, Proceedings of the SPE Annual Technical Conference and Exhibition, San Antonio, 5–8 Oct
82. Pereira GG, Pinczewski WV, Chan DYC, Paterson L, Øren PE (1996) Pore-scale network model for drainage-dominated three-phase flow in porous media. *Trans Porous Media* 24(2):167–201
83. Pike GE, Seager CH (1974) Percolation and conductivity: a computer study I. *Phys Rev B* 10(4):1421–1434
84. Piri M, Blunt MJ (2002) Pore-scale modeling of three-phase flow in mixed-wet systems. Paper SPE 77726, Proceedings of the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 29 September–2 October.
85. Prakash S, Havlin S, Schwartz M, Stanley HE (1992) Structural and long-range correlated percolation. *Phys Rev A* 46(4):R1724–R1727
86. Priest SD, Hudson JA (1981) Estimation of discontinuity spacings and trace length using scan line surveys. *Int J Rock Mech, Min Sci Geomech Abstr* 18:185–197
87. Renshaw CE (1999) Connectivity of joint networks with power law length distributions. *Water Resour Res* 35(9):2661–2670
88. Rives T, Razack M, Pett JP, Rawnsley KD (1992) Joint spacing: analogue and numerical simulation. *J Struct Geol* 14(8/9):925–937
89. Robinson PC (1983) Connectivity of fracture systems—a percolation theory approach. *J Phys A Math General* 16:605–614
90. Robinson PC (1984) Numerical calculations of critical densities for lines and planes. *J Phys A Math General* 17(14):2823–2830
91. Rossen WR, Gu Y, Lake LW (2000) Connectivity and permeability in fracture networks obeying power law statistics. Paper SPE 59720, Proceedings of the SPE Permian Basin Oil and Gas Recovery Conference, Midland, Texas, 21–23 March
92. Roslien J, King PR, Buldyrev S, Lopez E, Stanley HE (2004) Prediction oil production conditioned on breakthrough time. *Petrol Geosci* (submitted)
93. Rouleau A, Gale JE (1985) Statistical characterization of the fracture system in the Stripa Granite, Sweden. *International J Rock Mech, Min Sci Geomech Abstr* 22(6):353–367
94. Sahimi M (1994) Applications of percolation theory. Taylor and Francis, London
95. Sahimi M (1995) Flow and transport in porous media and fractured Rock. VCH publication, London, pp 103–157
96. Schmittbuhl J, Vilotte JP, Roux S (1993) Percolation through self affine surfaces. *J Phys A Math General* 26:6115–6133
97. Segall P, Pollard DD (1983) Joint formation in granitic rock in the Sierra Nevada. *Geol Soc Am Bull* 94:563–575
98. Snow DT (1969) Anisotropic permeability of fractured media. *Water Resour Res* 5(6):1273–1289
99. Soll WE, Celia MA (1993) A modified percolation approach to simulating three-fluid capillary pressure-saturation relationships. *Adv Water Resour* 16:107–126
100. Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor and Francis, London
101. Valvatne PH, Blunt MJ (2003) Predictive pore-scale network modeling. Paper SPE 84550, Proceedings of the SPE Annual Technical Conference and Exhibition, Denver, Colorado, 5–8 Oct
102. Van Dijk JP, Bello M, Toscano C, Bersani A, Nardon S (2000) Tectonic model and three-dimensional fracture analysis of Monte Alpi-Southern Apennines. *Tectonophysics* 324:203–237
103. Vermilye JM, Scholz CH (1995) Relationship between vein length and aperture. *J Struct Geol* 17(3):423–434
104. Watanabe K, Takahashi H (1995) Fractal geometry characterization of geothermal reservoir fracture networks. *J Geophys Res* 100(B1):521–528
105. Watanabe H, Yukawa S, Ito N, Hu C-K (2004) Superscaling of percolation on rectangular domains. *Phys Rev Lett* 93(19):190601.1–190601.4
106. Wilkinson D (1984) Percolation model of immiscible displacement in the presence of buoyancy forces. *Phys Rev A* 34(1):520–531
107. Wilkinson D, Willemsen JF (1983) Invasion percolation: a new form of percolation theory. *J Phys A Math General* 16(16):3365–3376
108. Xia W, Thrope MF (1988) Percolation properties of random ellipses. *Phys Rev A* 38(5):2650–2656
109. Zhang X, Sanderson DJ (2002) Numerical modelling and analysis of fluid flow and deformation of fractured rock masses. Elsevier Science, Pergamon

## Percolation Thresholds, Exact

JOHN C. WIERMAN

Department of Applied Mathematics and Statistics,  
Johns Hopkins University, Baltimore, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Trees

Two-Dimensional Bond Models

Site Models in Two Dimensions

Random Voronoi Percolation

Multiparameter Critical Surfaces

Future Directions

Bibliography

### Glossary

**Archimedean lattices** A *regular tiling* is a tiling of the plane which consists entirely of regular polygons. (A regular polygon is one in which all side lengths are equal and all interior angles are equal.) An *Archimedean lattice* is the graph of vertices and edges of a regular tiling which is vertex-transitive, i. e., for every pair of vertices,  $u$  and  $v$ , there is a graph isomorphism that maps  $u$  to  $v$ . There are exactly 11 Archimedean lattices. A notation for Archimedean lattices, which can also serve as a prescription for constructing them, is given in Grünbaum and Shephard [5]. Around any vertex (since all are equivalent, by vertex-transitivity), starting with the smallest polygon touching the vertex, list the number of edges of the successive polygons around the vertex. For convenience, an exponent is used to indicate that a number of successive polygons have the same size.

**Bond percolation** In a bond percolation model, a random subgraph is formed from an infinite graph  $G$  by retaining each edge of  $G$  with probability  $p$ , independently of all other edges.

**Dual graph** A graph is planar if it may be drawn in the plane with no edges intersecting except at their endpoints, thus dividing the plane into faces. Every planar graph  $G$  has a dual graph, denoted here by  $D(G)$ .  $D(G)$  may be constructed by placing a vertex of  $D(G)$  in each face of  $G$  and connecting two vertices of  $D(G)$  by an edge if the corresponding faces in  $G$  share a common edge. Note that  $D(D(G)) = G$ .

**Line graph** The line graph,  $L(G)$ , of a graph  $G$  is constructed by placing a vertex of  $L(G)$  on each edge of  $G$

and connecting two vertices of  $L(G)$  if the corresponding edges of  $G$  share a common endpoint.

**Matching graphs** A pair of matching graphs may be constructed from an underlying planar graph. Select a set  $F$  of faces of the graph. Construct a graph  $G$  by adding an edge in each face of  $F$  between any pair of vertices that are not already connected by an edge. Construct the matching graph  $M(G)$  of  $G$  by adding an edge between any pair of vertices in each face not in  $F$  that are not already connected by an edge. Note that  $M(M(G)) = G$ .

**Percolation threshold** In a percolation model with parameter  $p$ , there is a retention probability  $p_c$ , called the percolation threshold, above which the random subgraph contains an infinite connected component and below which all connected components are finite.

**Periodic graph** A periodic graph is an infinite graph that can be represented in  $d$ -dimensional space so that it is invariant under translations by all integer linear combinations of a fixed basis.

**Site percolation** In a site percolation model, a random graph is formed from an infinite graph  $G$  by retaining each vertex of  $G$  with probability  $p$ , independently of all other vertices. An edge of  $G$  is retained in the random graph if both its endpoint vertices are retained.

### Definition of the Subject

Percolation models were introduced in the 1950s by Broadbent and Hammersley [1] to model the flow of fluid in a random medium. Since both terms, fluid and medium, may be broadly interpreted, percolation has a wide variety of applications, including thermal phase transitions, oil flow in sandstone, and the spread of epidemics. An important motivation for the development of percolation models was to provide an alternative to diffusion models, in which the randomness was associated with the fluid while the medium is relatively homogeneous. Since percolation models associate the randomness with the medium, it is possible for the fluid either to become trapped or to flow infinitely far. This presence of a phase transition is an important reason for the importance of percolation models. The percolation threshold is a critical probability in the percolation model which corresponds to the phase transition point. Since the emphasis in percolation theory is on the effect of the medium on the behavior of the model, it is important to understand how the percolation threshold depends on the characteristics of the medium.

The medium is often, but not always, modeled by a periodic graph, representing an atomic lattice structure. In the bond percolation model, each edge of the graph is retained with probability  $p$ , independently of all other

edges, to create a random subgraph. In the site percolation model, each vertex of the graph is retained with probability  $p$ , independently of all other vertices, and each edge is retained if and only if both its endpoints are retained. In both models, the focus is on the properties – in particular, the size – of the connected components, called clusters, of the random subgraph.

The most common definition is that the percolation threshold is a retention probability value  $p_c$  such that if  $p > p_c$  there exists an infinite cluster in the graph and if  $p < p_c$  there are only finite clusters in the graph. However, there exist other interpretations, which correspond to different definitions of the percolation threshold: (1) A percolation threshold  $p_H$  is the critical probability above which a specific vertex  $v$  is in an infinite cluster with positive probability. The percolation threshold  $p_H$  is independent of the specific vertex  $v$  if the graph is connected. (2) A percolation threshold  $p_T$  is defined as the critical probability above which the expected cluster size containing a specific vertex  $v$  is infinite, and is also independent of the choice of  $v$  if the graph is connected. (3) For periodic graphs, a percolation threshold  $p_S$  may be defined in terms of the limiting behavior of the probability that a cluster connects opposite sides of a rectangle in a sequence of similar rectangles whose areas are increasing to infinity. For the periodic graphs discussed in this article, these definitions provide equal values for the percolation threshold.

Due to the dependence of the percolation threshold on the features of the lattice, since the origins of percolation theory much research has been devoted to deriving exact values, computing simulation estimates, and proposing approximation formulas for the percolation threshold as a function of the lattice. This article focuses on the extent of knowledge of exact values of percolation thresholds for bond percolation and site percolation on various graphs.

## Introduction

Although the percolation threshold problem is simply described and easily visualized, it has become recognized as extremely intractable, with the result that, after 50 years of research, exact percolation thresholds are known for few graphs. Besides the trivial one-dimensional case, and infinite regular trees, the only solutions are for two-dimensional graphs. There are no exact solutions for periodic graphs in three dimensions or higher.

We now provide a brief history of the development of exact percolation thresholds and the mathematical tools used.

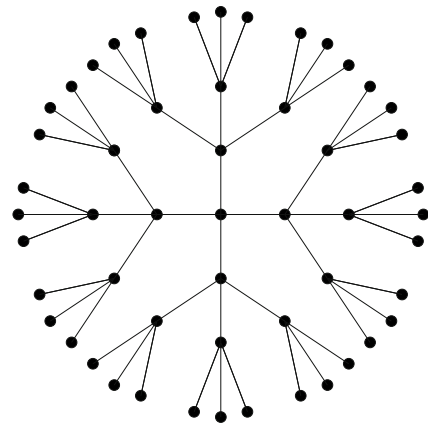
The first major development was in 1960 by Harris [6], who proved a lower bound of  $1/2$  for the square lattice

bond model threshold. The value was larger than simulation estimates at that time, and was believed to be sharp. Harris used the self-duality of the square lattice extensively in the proof, and established a lemma regarding covariant events which was later generalized by Fortuin, Kasteleyn, and Ginibre [4] and played a crucial role in later exact percolation threshold proofs.

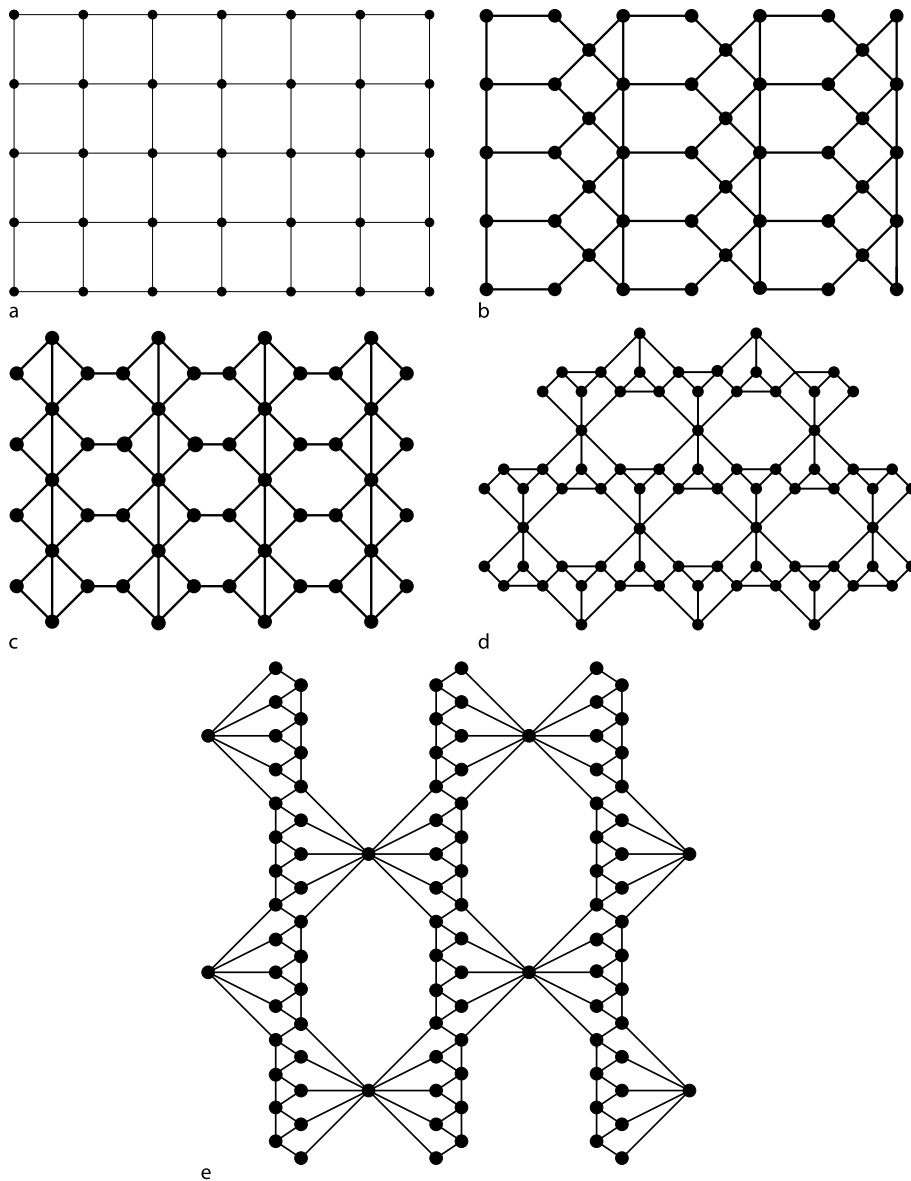
Recognizing the importance of duality in the study of bond percolation, Sykes and Essam [15] developed a corresponding concept of matching graphs for site percolation models. Interpreting the percolation threshold as a singularity of a clusters-per-site function, they derived values of the bond percolation thresholds for the square, triangular, and hexagonal lattices. Their methods implied that site percolation thresholds of matching graphs sum to one, and bond percolation thresholds of dual graphs sum to one. A key transformation in the solutions for the triangular and hexagonal lattices was a star-triangle transformation relating the two graphs. However, their exact values would not be given mathematically rigorous proofs until much later.

Although they did not establish the exact percolation thresholds, Seymour and Welsh [14], in the context of the square lattice bond model, laid important groundwork for the solution that followed. They recognized and defined the  $p_H$ ,  $p_T$ , and  $p_S$  interpretations of the percolation threshold, and proved relationships among them. Russo [11,12] independently established similar results for the square lattice site model.

In 1980, Kesten [7] rigorously established that the percolation threshold of the square lattice bond model is  $1/2$ , using the self-duality of the square lattice, and proving that all versions of the percolation threshold are equal in this case.



**Percolation Thresholds, Exact, Figure 1**  
A portion of a Cayley tree with vertex degree four



Percolation Thresholds, Exact, Figure 2

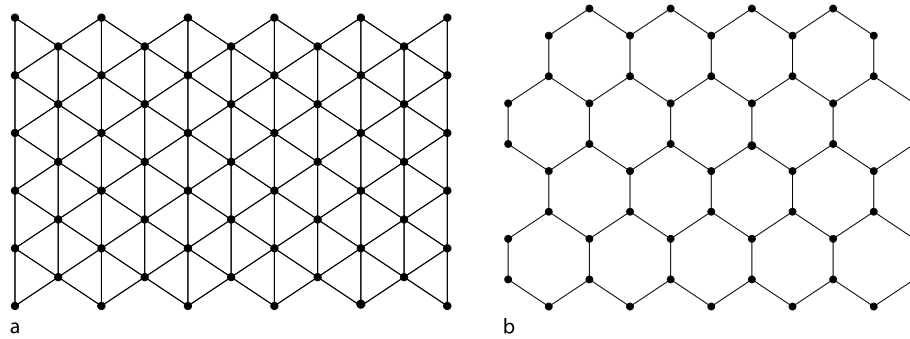
Portions of five self-dual periodic graphs. The square lattice is at the *upper left*. An example of the generalization to a family of self-dual graphs is at the *bottom*

Using Kesten's methods, the duality of the triangular and hexagonal lattices, and the star-triangle transformation between them, in 1981 Wierman [17] proved that the bond percolation threshold of the triangular lattice is the root of  $1 - 3p + p^3$  in the interval  $[0, 1]$ , which is equal to  $2 \sin(\pi/18) \approx .347296$ , and the bond percolation threshold of the hexagonal lattice is the complementary value, approximately .652704. In 1984, Wierman [19] discovered another pair of dual lattices for which the ex-

act bond percolation threshold could be determined using a version of the star-triangle transformation. The bond threshold of the bowtie lattice is the root of the polynomial  $1 - p - 6p^2 + 6p^3 - p^5$  in  $[0, 1]$ , which is approximately .404518, while the bond threshold of its dual graph is the complementary value, approximately .595482.

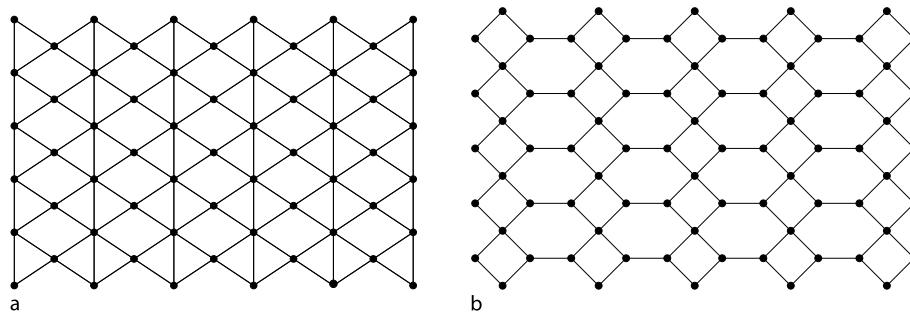
These results were generalized by Kesten [8] in his 1982 monograph, where he proved that the site percolation thresholds of a pair of periodic matching graphs





Percolation Thresholds, Exact, Figure 3

Portions of the triangular (a) and hexagonal or honeycomb (b) lattices



Percolation Thresholds, Exact, Figure 4

Portions of the bow-tie lattice (a) and its dual lattice (b)

sum to one. Since it is fully-triangulated and therefore self-matching, the triangular lattice has percolation threshold equal to one-half. The duality result for bond percolation thresholds is implied by this result via the bond-to-site transformation.

In 2006, Scullard and Ziff [13,21,22] derived exact bond percolation thresholds of additional periodic lattices based on the star-triangle transformation.

Exact percolation thresholds can be derived for additional graphs that are obtained by various transformations of graphs with exact solutions. Such solutions can be established via the bond-to-site transformation, subdivision of edges, and replacing edges with more complex decorations.

## Trees

A tree is a graph which is connected and has no cycles, or equivalently, which has a unique path between every pair of vertices. An infinite tree in which every vertex has the same degree is called a Cayley tree or Bethe lattice.

The earliest non-trivial exact percolation threshold solutions were for Cayley trees. Let  $C_k$  denote the Cayley tree

of degree  $k$ . Then, for all  $k \geq 3$ ,

$$p_c(C_k) = \frac{1}{k-1},$$

for both bond percolation and site percolation.

It is easy to see that the bond and site percolation thresholds are equal. In a bond model on a Cayley tree, consider starting from a specific vertex, called the root, and moving outward. For each edge, consider the vertex at the end farthest from the root to be retained or not according to whether the edge is retained or not. This creates a site percolation model with the same parameter value as the original bond percolation model, in which there is an infinite cluster if and only if there is an infinite cluster in the bond model. Thus, the percolation thresholds of the two models are equal.

The values of the thresholds for Cayley trees can be determined either by calculation of the probability that a vertex is in an infinite component or by elementary theory of branching processes.

Lyons [9] has shown that, for rooted trees in general, an average number of branches per vertex, called the branching number, may be defined. The percolation

threshold of the tree is the reciprocal of the branching number. However, the definition of the branching number is rather intricate, so exact percolation thresholds are not easily computed.

## Two-Dimensional Bond Models

Planarity plays an important role in establishing exact bond percolation thresholds in two-dimensional models. A graph is planar if it may be drawn in the plane with no edges intersecting except at their endpoints, thus dividing the plane into faces. Every planar graph  $G$  has a dual graph, denoted here by  $D(G)$ .  $D(G)$  may be constructed by placing a vertex of  $D(G)$  in each face of  $G$  and connecting two vertices of  $D(G)$  by an edge if the corresponding faces in  $G$  share a common edge. Note that  $D(D(G)) = G$ .

The computational importance of dual graphs is that the bond percolation thresholds of a pair of periodic matching graphs sum to one, as implied via the bond-to-site transformation by a result Kesten in 1982.

Without an additional relationship between the pair of dual graphs, duality itself does not yield an exact percolation threshold solution. However, if the two graphs are isomorphic, the common graph is said to be self-dual, and its bond threshold must be one-half. Thus, the bond percolation threshold of the square lattice is exactly one-half.

Besides the square lattice, there are other periodic self-dual graphs, which, for the same reason, have a bond percolation threshold equal to one-half. Figure 2 shows additional self-dual graphs, and illustrates a construction of an infinite family of periodic self-dual graphs, given in [20].

Essentially all other exact bond threshold solutions are derived using a relationship called the star-triangle transformation, first used by Sykes and Essam. Notice that the set of edges of the triangular lattice may be decomposed into triangles that are similarly oriented. If each triangle is replaced by a three-pointed star with the points at the vertices of the triangle, the resulting graph is the hexagonal lattice, which is the dual graph of the triangular lattice. If retention probability parameters of the two lattices can be found so that the probabilities of all possible events involving connections of the three vertices on the boundary of the triangle are equal, then the exact percolation threshold can be determined. The solution is the root of a polynomial involving the retention probability of the triangular lattice:  $1 - 3p + p^3$ , giving the solution  $2 \sin(\pi/18) \approx .347296$ . By duality, the bond threshold for the hexagonal lattice is  $1 - 2 \sin(\pi/18) \approx .652704$ . While this solution was derived by Sykes and Essam in 1964, mathematical methods to rigorously prove it were not de-

veloped until later, with the result being proved by Wierman in 1981.

Modified versions of the star-triangle transformation were used to derive other exact bond percolation thresholds. Wierman discovered a pair of lattices, called the bow-tie lattice and its dual, which could be solved exactly. The bond threshold of the bow-tie lattice is the root of  $1 - p - 6p^2 + 6p^3 - p^5$ , which is approximately .404518, while the dual lattice has threshold approximately .595482.

Scullard and Ziff [13,21,22] used a modified star-triangle approach to find values for a lattice that they named the martini lattice, and applied the approach to other planar two-dimensional graphs. They use a triangle-triangle transformation in the derivation of their results. While their approach does produce correct exact percolation threshold results for some graphs, further study is needed to determine the complete range of validity of their method.

For any of the exact bond threshold solutions, additional exact thresholds may be determined for certain transformations of the graphs.

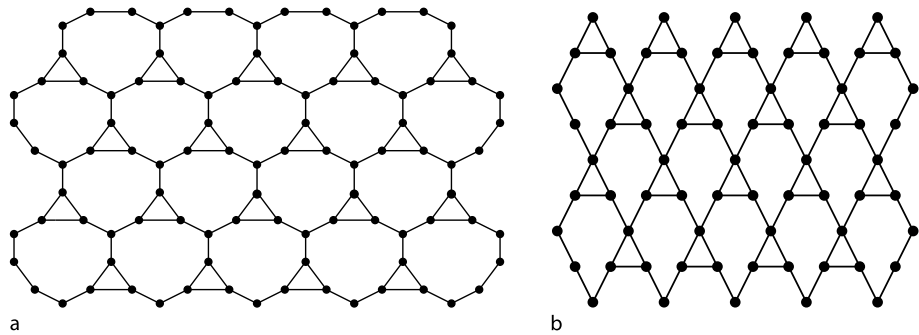
If each edge of a graph  $G$  is replaced by  $k$  edges in series, the resulting graph is called a  $k$ -subdivision of  $G$ . Since the  $k$  edges in series play the role of one edge of  $G$ , the bond percolation threshold of a  $k$ -subdivision of  $G$  is the  $k$ th root of the bond threshold of  $G$ .

More generally, instead of replacing each edge by a series of edges, one may replace it by some finite graph connecting only the two endpoints, which is called a decoration by Ord and Whittington [10]. By calculating the edge retention probability that makes the probability of connection through the decoration equal to the threshold of the original graph, the percolation threshold of the decorated graph may be exactly determined.

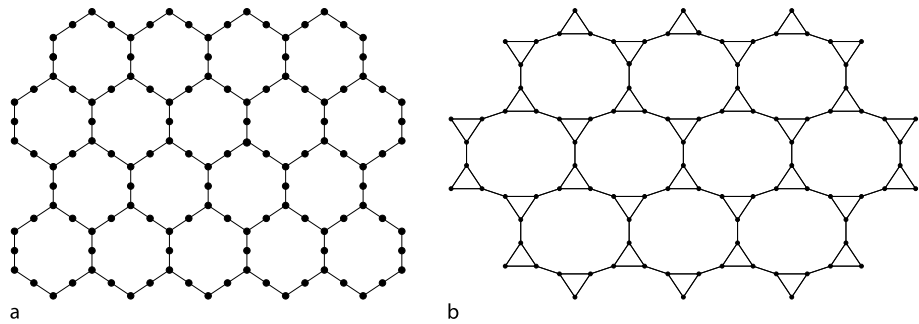
## Site Models in Two Dimensions

A key concept for the understanding of exact percolation threshold solutions for two-dimensional site models is the idea of a matching pair of graphs, introduced by Sykes and Essam in 1964.

A pair of matching graphs is constructed from an underlying planar graph. Select a set  $F$  of faces of the graph. Construct a graph  $G$  by adding an edge between any pair of vertices in each face of  $F$  that are not already connected by an edge. Construct the matching graph  $M(G)$  of  $G$  by adding an edge between any pair of vertices in each face not in  $F$  that are not already connected by an edge. Note that  $M(M(G)) = G$ . Note also that if the underlying planar graph has faces with more than three sides, then at least one of the graphs in the matching pair is nonplanar.



Percolation Thresholds, Exact, Figure 5  
Portions of the martini lattice (a) and martini-A lattice (b)



Percolation Thresholds, Exact, Figure 6  
Portions of the 2-subdivision of the hexagonal lattice (a) and its line graph, the  $(3, 12^2)$  lattice (b)

Percolation Thresholds, Exact, Table 1  
Exact bond percolation thresholds of selected lattice graphs

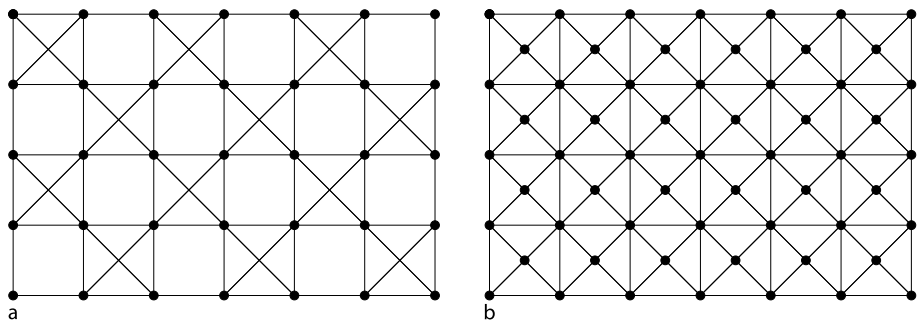
Lattice	Bond Threshold	Equation
Triangular = $(3^6)$	$2 \sin \pi/18 \approx .347296$	$p^3 - 3p + 1 = 0$
Bow-tie	.414518	$1 - p - 6p^2 + 6p^3 - p^5 = 0$
Square = $(4^4)$	.500000	$2p - 1 = 0$
Self-dual	.500000	$2p - 1 = 0$
D(Bow-tie)	.595482	$1 - p_c(\text{Bow-tie})$
Martini-A	.625457	$p^5 - 4p^4 + 3p^3 + 2p^2 - 1 = 0$
Hexagonal = $(6^3)$	.652704	$1 - p_c(3^6)$
Martini	$1/\sqrt{2} \approx .707107$	$(2p^2 - 1)(p^4 - 3p^3 + 2p^2 + 1) = 0$

The importance of matching graphs is that the site percolation thresholds of a pair of periodic matching graphs sum to one, as proved by Kesten [8] in 1982. In fact, this result implied the result for bond percolation thresholds for dual graphs, via the bond-to-site transformation.

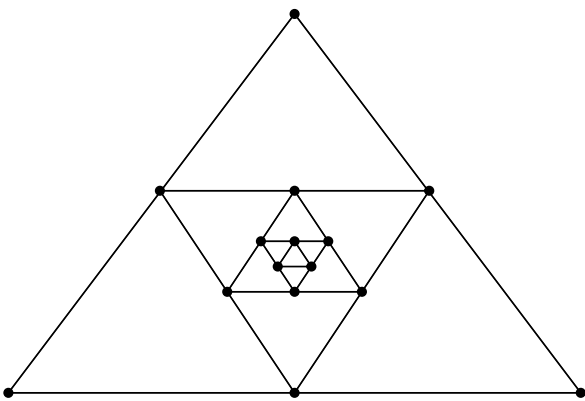
In general, additional information besides the matching property is needed to identify the exact percolation thresholds of the pair of graphs. However, if the matching graphs are identical, the graph is called self-matching, and the percolation threshold is necessarily one-half. Note that if a planar graph has all triangular faces, then it is self-

matching. Therefore, the site percolation threshold is exactly one-half for the triangular lattice and the dual graphs of the  $(4, 8^2)$ ,  $(4, 6, 12)$ , and  $(3, 12^2)$  lattices. In addition, there are self-matching graphs that are not fully-triangulated, such as the line graph of the square lattice, shown in Fig. 7.

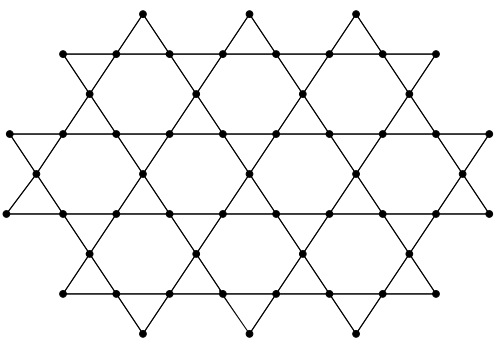
As a caution, however, note that a fully-triangulated graph may not have its site percolation threshold equal to one-half if it is not a periodic graph. An example of a fully-triangulated graph with site percolation threshold equal to one was given by van den Berg [16], and further discussion



Percolation Thresholds, Exact, Figure 7  
Two self-matching lattices: the line graph of the square lattice (a) and the dual of the  $(4, 8^2)$  lattice (b)



Percolation Thresholds, Exact, Figure 8  
A portion of van den Berg's counterexample: The graph is fully-triangulated, and thus self-matching, but has percolation threshold equal to one – not one-half



Percolation Thresholds, Exact, Figure 9  
A portion of the Kagomé lattice, which is the line graph of the hexagonal lattice

of similar counterexamples is provided in Wierman [18].  
Scullard and Ziff [13,21,22] have also proposed exact site percolation thresholds for additional two-dimensional lattices.

Additional exact site percolation threshold solutions have been obtained by transformations of bond models with exact solutions. The line graph,  $L(G)$ , of a graph  $G$  is constructed by placing a vertex of  $L(G)$  on each edge of  $G$  and connecting two vertices of  $L(G)$  if the corresponding edges of  $G$  share a common endpoint. If there is a bond percolation model on  $G$  with each edge retained with probability  $p$  independently of the other edges, one may define a site percolation model on  $L(G)$  in which each vertex is retained if and only if the corresponding edge of  $G$  is retained. Then, an infinite cluster in the bond model on  $G$  corresponds to an infinite cluster in the site model on  $L(G)$ , so the percolation thresholds of the two models are equal. This construction and relationship is called

Percolation Thresholds, Exact, Table 2  
Exact bond percolation thresholds of selected lattice graphs

Lattice	Site Threshold	Equation
Triangular = $(3^6)$	.500000	$2p - 1 = 0$
Self-matching	.500000	$2p - 1 = 0$
Kagomé = $(3, 6, 3, 6)$	.652704	$p = p_c(6^3 \text{ bond})$
$(3, 12^2)$	.807901	$p = \sqrt{p_c(6^3 \text{ bond})}$

the bond-to-site transformation, and allows translation of all exact bond percolation threshold solutions into exact site percolation threshold solutions on line graphs. For example, the line graph of the hexagonal lattice, called the Kagomé lattice, has a site percolation threshold of exactly  $1 - 2 \sin(\pi/18) \approx .652704$ , while its matching graph, the line graph of the triangular lattice, has exact site percolation threshold  $2 \sin(\pi/18) \approx .347296$ . As another example, the site percolation threshold of the  $(3, 12^2)$  lattice is exactly  $\sqrt{1 - 2 \sin(\pi/18)} \approx .807901$ , since it is the line graph of the 2-subdivision of the hexagonal lattice.

### Random Voronoi Percolation

In 2006, Bollobás and Riordan [2] provided the first exact percolation threshold solution for a continuum percolation model.

Consider the set of points  $P$  in a two-dimensional homogeneous Poisson point process. For a point  $p \in P$ , the Voronoi polygon of  $p$  is the set of all points in the plane that are closer to  $p$  than to any other point in  $P$ . Each Voronoi polygon is a convex polygon, and two Voronoi polygons either intersect in an edge or not at all. The edges of the collection of Voronoi polygons form an infinite planar graph called the Voronoi tessellation corresponding to  $P$ .

The dual of the Voronoi tessellation is called the Delaunay triangulation, since with probability one all faces are triangles. The remarkable result of Bollobás and Riordan [2] is that the site percolation threshold of the Delaunay triangulation is exactly one-half.

### Multiparameter Critical Surfaces

For some applications, multi-parameter percolation models are considered. For example, in a bond percolation model, edges in different directions may have different retention probability parameters, giving a multi-dimensional parameter space. In such a parameter space, the role of the percolation threshold is played by the boundary between regions of the parameter space where infinite clusters occur and where all clusters are finite, called the critical surface.

Some multi-parameter bond percolation models are exactly solved. Two notable examples are: (1) In the square lattice with vertical edges retained with probability  $p$  and horizontal edges retained with probability  $q$ , the critical surface is the line segment  $p + q = 1$ . (2) For the triangular lattice with retention probabilities  $r$ ,  $s$ , and  $t$  for bonds in the three different directions, the critical surface is the surface given by  $1 - r - s - t + rst = 0$ .

### Future Directions

As seen above, the exact bond or site percolation threshold is known for relatively few lattices, with the exact solutions restricted to infinite trees and two-dimensional periodic graphs. Although they have been studied extensively, exact thresholds are not known for such common graphs as the square lattice site model, the hexagonal lattice site model, and the Kagomé lattice bond model. Although there are tools such as duality and matching, there is no general method available for providing exact threshold values. The

grand challenge is to find the exact bond or site percolation threshold for a lattice in three dimensions or higher.

### Bibliography

#### Primary Literature

1. Broadbent SR, Hammersley JM (1957) Percolation processes. I. Crystals and mazes. *Proc Camb Philos Soc* 53:629–641
2. Bollobás B, Riordan O (2006) The critical probability for random Voronoi percolation in the plane is  $\frac{1}{2}$ . *Probab Theory Relat Fields* 136:417–468
3. Bollobás B, Riordan O (2006) A short proof of the Harris-Kesten Theorem. *Bull Lond Math Soc* 38:470–484
4. Fortuin C, Kasteleyn PW, Ginibre J (1971) Correlation inequalities on some partially ordered sets. *Commun Math Phys* 22: 89–103
5. Grünbaum B, Shephard GC (1987) *Tilings and Patterns*. WH Freeman, New York
6. Harris TE (1960) A lower bound for the critical probability in a certain percolation process. *Proc Camb Philos Soc* 56:13–20
7. Kesten H (1980) The critical probability of bond percolation on the square lattice equals  $\frac{1}{2}$ . *Commun Math Phys* 74:41–59
8. Kesten H (1982) *Percolation Theory for Mathematicians*. Birkhäuser, Boston
9. Lyons R (1990) Random walks and percolation on trees. *Ann Probab* 18:931–958
10. Ord G, Whittington SG (1980) Lattice decorations and pseudo-continuum percolation. *J Phys A: Math Gen* 13:L307–L310
11. Russo L (1978) A note on percolation. *Z Wahrscheinlichkeitstheorie Verwandte Geb* 43:39–48
12. Russo L (1981) On the critical percolation probabilities. *Z Wahrscheinlichkeitstheorie Verwandte Geb* 56:229–237
13. Scullard CR (2006) Exact site percolation thresholds using a site-to-bond transformation and the star-triangle transformation. *Phys Rev E* 73:016107
14. Seymour PD, Welsh DJA (1978) Percolation probabilities on the square lattice. *Ann Discret Math* 3:227–245
15. Sykes MF, Essam JW (1964) Exact critical percolation probabilities for site and bond problems in two dimensions. *J Math Phys* 5:1117–1127
16. van den Berg J (1981) Percolation theory on pairs of matching lattices. *J Math Phys* 22:152–157
17. Wierman JC (1981) Bond percolation on the honeycomb and triangular lattices. *Adv Appl Probab* 13:298–313
18. Wierman JC (1984) Counterexamples in percolation: the site percolation critical probabilities  $p_H$  and  $p_T$  are unequal for a class of fully triangulated graphs. *J Phys A: Math Gen* 17: 637–646
19. Wierman JC (1984) A bond percolation critical probability determination based on the star-triangle transformation. *J Phys A: Math Gen* 17:1525–1530
20. Wierman JC (2006) Construction of infinite self-dual graphs. *Proceedings of the 5th Hawaii International Conference on Statistics, Mathematics and Related Fields (CD-ROM)*. East West Council for Education, Honolulu
21. Ziff RM (2006) Generalized cell – dual-cell transformation and exact thresholds for percolation. *Phys Rev E* 73:016134
22. Ziff RM, Scullard CR (2006) Exact bond percolation thresholds in two dimensions. *J Phys A: Math Gen* 39:15083–15090



## Books and Reviews

- Bollobás B, Riordan O (2006) Percolation. Cambridge University Press, Cambridge
- Grimmett G (1999) Percolation, 2nd edn. Springer, Berlin
- Hughes BD (1996) Random Walks and Random Environments, vol 2: Random Environments. Oxford Science Publications, Oxford

## Periodic Orbits of Hamiltonian Systems

LUCA SBANO

Mathematics Institute, University of Warwick,  
Warwick, UK

### Article Outline

Glossary  
Definition  
Introduction  
Periodic Solutions  
Poincaré Map and Floquet Operator  
Hamiltonian Systems with Symmetries  
The Variational Principles and Periodic Orbits  
Further Directions  
Acknowledgments  
Bibliography

### Glossary

**Hamiltonian** are called all those dynamical systems whose equations of motion form a vector field  $X_H$  defined on a symplectic manifold  $(\mathcal{P}, \omega)$ , and  $X_H$  is given by  $i_{X_H}\omega = dH$ , where  $H: \mathcal{P} \rightarrow \mathbb{R}$  is the Hamiltonian function.

**Poisson systems** These are dynamical systems whose vector field  $X_H$  can be described through a Poisson structure (Poisson brackets) defined on the ring of differentiable functions on a given manifold that is not necessarily symplectic (see “[Hamiltonian Equations](#)”). Note that on any symplectic manifold there is a natural Poisson structure such that any Hamiltonian system admits a Poisson formulation, but the contrary is false. The Poisson formulation of the dynamics is a generalization of the Hamiltonian one.

**A periodic orbit**  $\phi(\cdot)$  is a solution of the equations of motion that repeats itself after a certain time  $T > 0$  called a *period*, that is,  $\phi(t + T) = \phi(t)$  for every  $t$ .

**Poincaré section/map** Given a periodic orbit  $\phi(\cdot)$  a Poincaré section is a hyperplane  $S$  intersecting the curve  $\{\phi(t): t \in [0, T)\}$  transversely. The associated

Poincaré map  $\Pi$  maps neighborhoods of  $S$  into itself by following the orbit  $\phi(\cdot)$  (see Definition 10).

**A Hamiltonian system with symmetry** is a Hamiltonian system in which there is a group  $G$  acting on  $\mathcal{P}$ , i.e., there is a map  $\Phi: G \times \mathcal{P} \mapsto \mathcal{P}$ , with  $\Phi$  preserving the Hamiltonian and the symplectic form.

**Relative periodic orbit** Let  $G$  be a symmetry group for the dynamics. A path  $\phi(\cdot)$  is a relative periodic orbit if solves the equations of motion and repeats itself up to a group action after a certain time  $T > 0$ , that is,  $\phi(t + T) = \Phi_g(\phi(t))$  for every  $t$  and for some  $g \in G$ .

**Continuation** Continuation is a procedure based on the implicit function theorem (IFT) that allows one to extend the solution of an equation for different values of the parameters. Let  $f(x, \epsilon) = 0$  be an equation in  $x \in \mathbb{R}^n$  where  $f$  is differentiable and  $\epsilon \geq 0$  a parameter. Assume that  $f(x_0, 0) = 0$ ; a curve  $x(\epsilon)$  is called a *continued solution* if  $x(0) = x_0$  and  $f(x(\epsilon), \epsilon) = 0$  for some  $\epsilon \geq 0$ . In general  $x(\epsilon)$  exists whenever the IFT can be applied, that is, if  $D_x f(x, \epsilon)$  is invertible at  $(x_0, 0)$ .

**Liapunov–Schmidt reduction** Let  $f$  be a function on a Banach space. Liapunov–Schmidt reduction is a procedure that allows one to study  $f(x, \epsilon) = 0$  under the condition that the kernel of  $Df$  is not empty but it is finite-dimensional.

**Variational principles** The principles which aim to translate the problem of solving the equations of motion of a dynamical system (e.g., Hamiltonian systems) into the problem of finding the critical points of certain functionals defined on spaces of all possible trajectories of the given system.

### Definition

The study of periodic motions is very important in the investigations of natural phenomena. In particular the Hamiltonian formulation of the laws of motion has been able to formalize and solve many fundamental problems in mechanics and dynamical systems. This paper is focused on a selection of results in the study of periodic motions in Hamiltonian systems. We shall consider local problems (e.g., stability and continuation/bifurcation) and also the application of variational methods to study the existence in the large. Throughout the paper we give some details of the methods and proofs.

### Introduction

Periodic motions and behaviors in Nature have always been of interest to mankind. All phenomena that have some cyclic nature have captured our attention because

they are a sign and a clue for regularity. Therefore, they are indications of the possibility of understanding the laws of Nature. Since the second century BC, Greek philosophers and astronomers have looked into the possibility of describing the motions of celestial bodies through the theory of epicycles, which are combinations of circular periodic motions. Notably from a modern point of view this theory can be interpreted as a clever geometrical application of Fourier expansions of the observed motions [23]. The development of mechanics, the discovery that the laws of Nature can be written in the language of calculus and that laws of motion can be described in terms of differential equations opened up the study of periodic solutions of equations of motion. In particular, since Newton and then Poincaré [46], the main interest has been the understanding of the planetary motions and the solution of the so-called *N-body problem*, the reader should refer to ► [n-Body Problem and Choreographies](#) in this encyclopedia. The interest in periodic motions has not been restricted to celestial mechanics but became a sort of paradigm in all areas where mechanics was successfully applied. In this article we shall illustrate some general aspects of the results regarding the theory of periodic orbits in Hamiltonian systems. Such systems are the modern formulation of those mechanical systems which are described by second-order differential equations and have an energy function. As an example the reader could think of Newton's equations for a point mass in potential field. The equations read

$$m \ddot{x}(t) = -\nabla V(x(t)), \quad \text{with } x(t) \in \mathbb{R}^3 \quad (1)$$

for every  $t$ ,  $m$  is the mass

$V(x)$  is the potential and  $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$ . The energy function

$$E = \frac{m}{2} \|\dot{x}(t)\|^2 + V(x(t))$$

is conserved along the trajectories solving (1). It is important to say that most of the systems of interest in physics can be naturally written in Hamiltonian form. The plan of this article is as follows. First we introduce the Hamiltonian formulations of the equations of motion for a classical mechanical system. It is well known that in many applications Hamiltonian systems derive from a Lagrangian formulation; therefore, this is also presented. Furthermore we introduce the Poisson formulation that is essentially the first generalization to the Hamiltonian point of view. Then we turn to the study of the properties of periodic solutions. In particular we focus on their "local properties", persistence and stability. In this analysis the main tool will

be the implicit function theorem (IFT). In order to emphasize the utility of the IFT we present some of the proofs that contain typical calculations often scattered in the literature. Then we consider the problem of periodic orbit for Hamiltonian systems with symmetries, where we introduce the notion of relative equilibrium and relative periodic orbit. Inevitably we have also a short excursion about symmetry reduction, which is the natural theoretical setting to study systems with symmetries. The second part of the article is devoted to the exposition of the study of periodic orbits by variational methods. The discovery that the equations of motion of mechanical systems can be derived by a variational principle, the so-called *least action principle*, is usually attributed to Maupertuis (eighteenth century). According to this principle, the motions are critical points of a functional called the *action* defined in a suitable space of paths. Variational methods turned out to be one of the most effective methods to prove the existence of periodic orbits; notable is the case of the *N-body problem* (see [2] and ► [n-Body Problem and Choreographies](#)). The valuable feature of the variational methods is the possibility to study the existence problem by looking at the topology and geometry of the space of periodic paths without further restrictions. In the presentation of the results some elements of the proofs are illustrated in order to clarify the main ideas. In the final section there are some open problems and further directions of investigation, in particular a simple example of the so-called *multisymplectic* structures that has extended the possibility of applying the finite-dimensional Hamiltonian approach to multiperiodic problems for a large class of partial differential equations is presented. For centuries the study of periodic orbits has been one of the main centers of mathematical investigations and developments and still presents challenges and the capacity for producing new interesting mathematical ideas to understand the complexity of Nature.

## Hamiltonian Equations

A Hamiltonian system is given by specifying a symplectic manifold  $(\mathcal{P}, \omega)$ , where  $\mathcal{P}$  is a differentiable manifold of even dimension,  $\omega$  is a closed differential two-form and a function  $H: \mathcal{P} \rightarrow \mathbb{R}$  is called a Hamiltonian. In the language of the differential forms the Hamiltonian vector  $X_H$  field on  $\mathcal{P}$  is written as

$$i_{X_H} \omega = dH. \quad (2)$$

In the case  $\mathcal{P} = \mathbb{R}^{2n}$ , the symplectic structure is  $\omega_0 = \sum_{i=1}^n dx_i \wedge dy_i$  and the Hamiltonian vector field is

$$X_H(z) \doteq J \nabla_z H(z), \quad (3)$$

where  $z = (x, y) \in \mathbb{R}^{2n}$  and  $J$  is the symplectic matrix

$$J = \begin{pmatrix} 0 & \mathbf{id}_n \\ -\mathbf{id}_n & 0 \end{pmatrix}. \quad (4)$$

The Hamiltonian equations are then

$$\frac{dz(t)}{dt} = X_H(z(t)). \quad (5)$$

In the case  $\mathcal{P} = \mathbb{R}^{2n}$  (5) reads

$$\dot{z}(t) = (\dot{x}(t), \dot{y}(t)) = (\nabla_y H(x(t), y(t)), -\nabla_x H(x(t), y(t))).$$

For a mechanical systems described by (1) the Hamiltonian function is

$$H = \frac{\|y\|^2}{2m} + V(x), \quad \text{where } (x, y) \in \mathbb{R}^6,$$

and its Hamiltonian equations read

$$\begin{cases} \dot{x}(t) = \frac{y(t)}{m} \\ \dot{y}(t) = -\nabla V(x(t)). \end{cases}$$

Note that the first equation corresponds to the classical definition of *momentum* in mechanics (here denoted with  $y$ ) and the Hamiltonian function  $H$  coincides with the energy  $E$ . For more details the reader could consult [1,5] and also ► [Dynamics of Hamiltonian Systems](#) in this encyclopedia.

**Lagrangian Formulation** In many applications in physics and in particular in mechanics Hamiltonian systems arise from the Lagrangian description. In such a setting a mechanical system is described by prescribing a differentiable manifold  $\mathcal{M}$  (the *configuration space*) and a Lagrangian function  $L$  defined on the tangent bundle  $T\mathcal{M}$ . Let  $L: T\mathcal{M} \rightarrow \mathbb{R}$  be a Lagrangian on a manifold  $\mathcal{M}$  of dimension  $n$ . If  $L$  is hyperregular (i.e.,  $\text{rank}(D_{v_q}^2 L(v_q, q)) = n$ ), then the Hamiltonian function is naturally constructed on the cotangent bundle  $T^*\mathcal{M}$  by using the Legendre transform [1,5] as follows:

$$p_i = \frac{\partial L}{\partial v_q^i}, \quad (6)$$

$$H(p, q) = \sum_{i=1}^n v_q^i(p) q_i - L(v_q(p, q), q).$$

The Hamiltonian system is then defined on the cotangent bundle of  $\mathcal{M}$  that is  $\mathcal{P} = T^*\mathcal{M}$ , which is endowed with the canonical symplectic form  $\omega = d\theta$ , where

$\theta = \sum_{i=1}^n p_i dq_i$ . In the Lagrangian description the equations of motion are

$$\frac{d}{dt} \frac{\partial L(v_q(t), q(t))}{\partial v_{q_i}} - \frac{\partial L(v_q(t), q(t))}{\partial q_i} = 0, \quad (7)$$

$i = 1, \dots, n$  where  $v_{q_i}(t) = \dot{q}_i(t)$ ,  $i = 1, \dots, n$ .

Note that (7) contains second order time-derivatives. For more details see [1,5].

**Poisson Formulation** Let  $\mathcal{F}(\mathcal{P})$  be the space of differentiable functions on  $(\mathcal{P}, \omega)$ . On  $\mathcal{F}(\mathcal{P})$  can be introduced a product  $\{.,.\}$  [1,5,16]. The Poisson brackets

$$\omega(X_f, X_g) = \{f, g\} \quad \text{for } f, g \in \mathcal{F}(\mathcal{P}). \quad (8)$$

In terms of the Poisson brackets the Hamiltonian equations can be written as a derivation acting on  $\mathcal{F}(\mathcal{P})$ :

$$X_H(f) = \{f, H\} \quad \text{for } f \in \mathcal{F}(\mathcal{P}). \quad (9)$$

The Poisson brackets satisfy the following properties. For all  $f, g \in \mathcal{F}(\mathcal{P})$

$$\begin{aligned} \{f, g\} &\text{ is bilinear with respect to } f \text{ and } g, \\ \{f, g\} &= -\{g, f\} \\ \{fg, h\} &= f\{g, h\} + g\{f, h\} \\ \{\{f, g\}, h\} &+ \{\{h, f\}, g\} + \{\{g, h\}, f\} = 0 \text{ Jacobi identity} \end{aligned} \quad (10)$$

An easy consequence of (9) and (10) is

**Proposition 1** *The Hamiltonian function  $H$  is a constant of motion.*

A manifold  $\mathcal{P}$  endowed with the brackets  $\{.,.\}$  is called a *Poisson manifold*. Any symplectic manifold is a Poisson manifold [1,27] but the contrary is false. In fact Poisson brackets on a symplectic manifold are always nondegenerate, namely, the condition  $\{k, f\} = 0$  for all  $f \in \mathcal{F}(\mathcal{P})$  implies that  $k$  is identically zero. In a general Poisson manifold there might exists nonvanishing  $k$ , which are then called *Casimir* functions. This is related to the fact that symplectic manifolds are always even-dimensional, whereas Poisson manifolds can be odd-dimensional. We can look at Poisson manifolds as a useful generalization of Hamiltonian systems; in fact in order to define Poisson brackets it is sufficient to have the ring of functions  $\mathcal{F}(\mathcal{P})$ . For a general and complete description of the Poisson structure the reader could see [1,5,16,27].

## Periodic Solutions

Given a Hamiltonian vector field  $X_H(z)$  on  $(P, \omega)$ , one can consider the following Cauchy problem:

$$\begin{cases} \frac{dz(t)}{dt} = X(z(t)) \\ z(0) = z_0. \end{cases} \quad (11)$$

Equation (11) is meant to be defined on a local chart in  $P$ .

**Definition 1** We call *flow* or *integral flow* the map  $t \mapsto \phi(t, z_0)$  where  $z(t) = \phi(t, z_0)$  solves problem (11).

Some simple consequences follow [36].

*Remark 1* If  $X_H$  is complete, then the flow is defined for all  $t \in \mathbb{R}$ .

*Remark 2* If the Hamiltonian vector field is autonomous, (that is, not explicitly dependent on time), then  $\phi(., z_0)$  satisfies the composition property  $\phi(t, \phi(s, z_0)) = \phi(t + s, z_0)$  and  $\phi(., .)$  is called Hamiltonian flow.

Let us now introduce the main object of this exposition.

**Definition 2** A flow  $\phi: \mathbb{R} \rightarrow P$ ,  $z(t) = \phi(t, z_0)$  is said to be a  $T$ -periodic solution of (11) if there exists  $T > 0$  such that  $\phi(t + T, z_0) = \phi(t, z_0)$  for all  $t$ .

*Remark 3* Note that if  $\phi(., z_0)$  is a  $T$ -periodic solution, then  $\phi(., z_0)$  is  $nT$ -periodic for any  $n \in \mathbb{N}$ . In fact from the definition  $\phi(T, z_0) = z_0$  and using Remark 2, one can iterate

$$\begin{aligned} \phi(t + nT, z_0) &= \phi(t + (n-1)T, \phi(T, z_0)) \\ &= \phi(t + (n-1)T, z_0) \end{aligned}$$

and find  $\phi(t + nT, z_0) = \phi(t, z_0)$ .

**Definition 3**  $T$  is called a minimal period of  $\phi(t, z_0)$  if  $T = \min_{\tau \in \mathbb{R}_+} \{\tau: \phi(t + \tau, z_0) = \phi(t, z_0) \text{ for all } t\}$ .

**Definition 4** A point  $z^* \in P$  such that  $J \nabla_z H(z^*) = 0$  is called an equilibrium solution.

Obviously any equilibrium solution can be seen as a periodic solution with  $T = 0$ . One can easily show

**Lemma 1**  $\phi(t, z_0)$  is periodic of period  $T$  if and only if  $\phi(T, z_0) = z_0$ .

## Stability of Periodic Orbits

Given a periodic orbit the first natural question is to study its stability. There are three possible stability criteria [1, 36]:

**Definition 5 (Liapunov-stable)** A periodic orbit  $\phi(., z_0)$  is Liapunov-stable if for all  $\epsilon > 0$  there is  $\delta(z_0, \epsilon)$  such that  $\|z_0 - z'\| \leq \delta(z_0, \epsilon)$  implies that  $\|\phi(t, z_0) - \phi(t, z')\| \leq \epsilon$  for all  $t \geq 0$ .

The previous definition is natural for a non-Hamiltonian system but in a Hamiltonian context it is very strong, since in the Hamiltonian systems periodic orbits are not isolated. It is useful though to compare Liapunov stability with the following weaker notions.

**Definition 6 (Spectrally stable)** A periodic orbit  $\phi(t, z_0)$  is spectrally stable if the eigenvalues of  $DX_H(z_0)$  lie all on the unit circle.

**Definition 7 (Linearly stable)** A periodic orbit  $\phi(t, z_0)$  is linearly stable if it is spectrally stable and  $DX_H(z_0)$  can be diagonalized.

One can show that spectral stability is implied by either linear stability or Liapunov stability. A natural notion of stability can be introduced by using the Poincaré map and will be presented in Sect. “Poincaré Map and Floquet Operator”. The following results describe the structure of linear Hamiltonian systems, namely, systems whose Hamiltonian function is  $H(z) = \frac{1}{2} \langle z, A z \rangle$  and the equations of motion read  $\dot{z}(t) = J A z(t)$ .

**Theorem 1 ([36])** Let  $A$  be time-independent. The characteristic polynomial of  $J A$  is even and if  $\lambda$  is an eigenvalue then so are  $-\lambda, \bar{\lambda}, -\bar{\lambda}$ .

A consequence of the previous result is that linear stability is equivalent to spectral stability for linear autonomous Hamiltonian systems.

Linear systems can depend on parameters; it is therefore interesting to have a notion of stability with respect to parametric changes.

**Definition 8** A linear stable Hamiltonian system  $H(z) = \frac{1}{2} \langle z, A z \rangle$  is said to be parametrically stable if for every symplectic matrix  $B$  such that  $\|A - B\| < \epsilon$  the system  $\dot{z} = J B z$  is linearly stable.

We finally give a characterization of linear Hamiltonian systems which are parametrically stable.

**Theorem 2 ([36])** If the Hamiltonian  $H$  is positive (or negative) definite or all the eigenvalues are simple, then  $A$  is parametrically stable.

## Continuation and Bifurcation of Equilibrium Solutions

Let  $H(z, \epsilon)$  be a Hamiltonian function depending on a parameter  $\epsilon \geq 0$ . Let

$$Z_0 = \{z^*: J \nabla H(z^*, 0) = 0\}$$

be the set of equilibrium positions. An interesting problem is to study how  $Z_0$  is modified when  $\epsilon > 0$ . The local properties of  $Z_0$  depend on the spectrum of  $J D^2 H(z^*, 0)$ . If  $z^* \in Z_0$  is such that  $J D^2 H(z^*, 0)$  has nonzero eigenvalues, then by the IFT there exists a curve  $z^*(\epsilon)$  of equilibrium points for  $\epsilon$  positive and sufficiently small. If  $J D^2 H(z^*, 0)$  has at least one zero eigenvalue then bifurcation can occur, but the interesting point is that bifurcations are possible also when  $J D^2 H(z^*, 0)$  is not degenerate. In fact in Hamiltonian systems generically the spectrum of the linearization of the vector field contains couples of complex conjugated eigenvalues [5]. In such a case there is a theorem due to Liapunov which shows the existence of periodic orbits – so-called *nonlinear normal modes* in a neighborhood of  $z^*$ . Let us now present Liapunov's theorem. This result describes how a nondegenerate equilibrium can be continued into a periodic orbit. The types of orbits are called *nonlinear normal modes*.

**Theorem 3 (Liapunov's center theorem [1,18])** *If the Hamiltonian system has a nondegenerate equilibrium at which the linearized vector field has eigenvalues  $\pm i\omega, \lambda_3, \dots, \lambda_n$  with  $\lambda_k/\omega \notin \mathbb{Z}$  then there exists a one-parameter family of periodic orbits emanating from  $z^*$ . The period tends to  $2\pi/\omega$  when the orbit radius tends to zero and the nontrivial multipliers tend to  $\exp(2\pi\lambda_k/\omega)$  with  $k = 3 \dots n$ .*

*Proof* Without loss of generality the nondegenerate equilibrium can be fixed at the origin. In a neighborhood of the origin the Hamiltonian vector field can be written as follows:

$$\dot{z}(t) = JA z(t) + r(z) ,$$

where  $\|r(z)\| = o(\|z\|)$ , that is,  $\|r(z)\|$  is infinitesimal with respect to  $\|z\|$ .

The spectrum of  $JA$  is  $\text{Spec}(JA) = \{\pm i\omega, \lambda_3, \dots, \lambda_n\}$ . Let  $y = \epsilon z$  with  $\epsilon \in [0, 1]$ , then

$$\dot{y}(t) = JA y(t) + r(y, \epsilon) ,$$

with  $r(y, \epsilon)$  infinitesimal for  $\epsilon \rightarrow 0$  uniformly for  $\|y\|$  bounded. For  $\epsilon = 0$  the system becomes

$$\dot{y}(t) = JA y(t) \quad (12)$$

and admits a periodic solution with period  $T = 2\pi/\omega$ :

$$y_0(t) = \exp(t JA) y_0, \quad JA y_0 = \nu_0 y_0 .$$

Equation (12) is linear, and thereby coincides with its linearization. The Floquet multipliers are therefore

$$(1, 1, \exp(2\pi \lambda_3/\omega), \dots, \exp(2\pi \lambda_n/\omega))$$

with  $\exp(2\pi \lambda_k/\omega) \neq 1$  by hypothesis. Now we look for a solution of

$$\left(\frac{d}{dt} - JA\right) y(t) = r(y(t), \epsilon) . \quad (13)$$

On the space  $C^1([0, T], \mathbb{R}^{2n})$  with periodic boundary condition the operator  $d/dt - JA$  has a two-dimensional kernel. Therefore, one could solve (13) by looking for a solution in the form

$$y(t, \epsilon) = \exp(t JA) y_0 + u(t) ,$$

where  $u(t) \in \text{rank}(d/dt - JA)$ . By the IFT one can show that  $\|u(t)\| = o(1)$  and therefore the solution can be continued for small  $\epsilon$  as

$$y(t, \epsilon) = \exp(t JA)(1 + o(1)) y_0 ,$$

that is,

$$z(t, \epsilon) = \epsilon \exp(t JA)(1 + o(1)) y_0 \text{ with } \lim_{\epsilon \rightarrow 0} \|z(t, \epsilon)\| = 0 .$$

□

**Remark 4** The analysis of the operator  $d/dt - JA$  used to prove the previous result is known as Liapunov–Schmidt reduction. In Sect. “Continuation of Periodic Orbits as Critical Points” we shall give an application of it in the study of critical points.

### Normal Form Analysis Near Equilibrium Points

In [55] Liapunov's theorem was generalized to cases where the condition  $\lambda/\omega \in \mathbb{Z}$  might hold. This corresponds to the so-called *resonance condition*.

**Definition 9 (Resonance)** The set of eigenvalues  $\{\omega_l\}_{l=1}^k$  of the linearization  $DX_H(z_0)$  are said to be resonant if  $\mathbb{R}(\omega_i) = 0$  for all  $i$  and there exist  $\{n_l\}_{l=1}^k \subset \mathbb{Z}$  such that

$$\sum_{l=1}^k \omega_l n_l = 0 .$$

In [55] it is shown that around  $z_0$  there are at least  $n$  periodic orbits. Because of the possible presence of resonance not all the orbits are a continuation of periodic orbits of the linearized vector field. In order to study this case one has to go beyond the linear approximation and analyze the Hamiltonian system in a neighborhood of the equilibrium taking into account the structure of terms of order higher than 2 in the canonical coordinates. To achieve this objective there is a general method, the *normal form theory*, to



expand the Hamiltonian function  $H$  using suitable coordinates in an  $\epsilon$ -neighborhood of  $z_0$ . A given Hamiltonian  $H$  can be expanded as

$$H(z) = H_0(z) + \sum_{m=1} \epsilon^m H_m(z), \quad (14)$$

where

$$\begin{aligned} H_0(z) &= \sum_{j=1} \frac{\omega_j}{2} \|z_j - z_{0,j}\|^2 \\ &= \sum_{j=1} \frac{\omega_j}{2} [(x_j - x_{0,j})^2 + (y_j - y_{0,j})^2] \end{aligned}$$

and  $H_m(z)$  are polynomials of degree  $m + 2$  in the canonical coordinates. Normal form theory allows us to classify the possible form of expansions (14) according to the resonance condition. This classification is also very important in the study of small perturbations of integrable Hamiltonian systems (*KAM theory*) where it is still the resonance condition that causes the main difficulties [5]. For a detailed account of normal form theory and its applications the reader could refer to [49,54] and to ► [Dynamics of Hamiltonian Systems](#) for an introduction. Once the Hamiltonian is in normal form, one can study the bifurcation occurring when the resonance holds. Particularly interesting is to understand what conditions the frequencies have to fulfill in order for a system to have a number of periodic orbits exceeding the estimation given in [55]. There have been further generalizations to the results given in [55] and the reader could refer to [40,41].

### Poincaré Map and Floquet Operator

To study a periodic orbit one could consider looking at it in a hyperplane which is transverse to its direction. This can be done locally and by constructing the *Poincaré map*.

**Definition 10** Let  $\phi(\cdot, z_0)$  be a  $T$ -periodic orbit. A Poincaré cross section of  $\phi(\cdot, z_0)$  is

$$S_a = \{z \in \mathcal{P} : \langle a, (z - z_0) \rangle = 0, \text{ with } \langle a, J\nabla H(z_0) \rangle \neq 0\}.$$

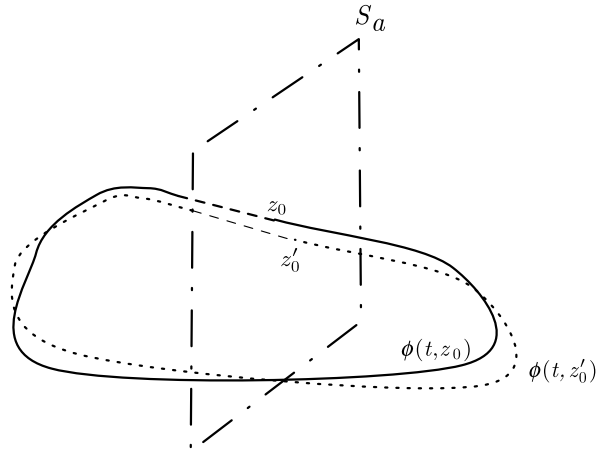
Let  $U \subset S_a$ , a neighborhood of  $z_0$  sufficiently small such that the first return time is

$$\mathcal{T} : U \rightarrow \mathbb{R}_+, \quad \phi(\mathcal{T}(z), z) \in S_a \text{ with } \mathcal{T}(z_0) = T.$$

The Poincaré map  $\Pi : U \rightarrow S_a$  is defined as

$$U \ni z \rightarrow \Pi(z) \doteq \phi(\mathcal{T}(z), z) \in S_a.$$

Using the regularity properties of  $\phi(t, z_0)$ , one can show



**Periodic Orbits of Hamiltonian Systems, Figure 1**

$S_a$  is Poincaré section for  $\phi(t, z_0)$ , in this case  $a = X_H(z_0)$ . The continuation produces a new orbit with initial data  $z'_0$  close to  $z_0$

**Proposition 2** ([36]) Let  $\sigma(t, z) \doteq \langle a, \phi(t, z) - z \rangle$  be defined for  $z \in S_a$ . For  $U \subset S_a$  sufficiently small the return times  $\mathcal{T} : U \rightarrow \mathbb{R}_+$  defined by the implicit equation  $\sigma(\mathcal{T}(z), z) = 0$  and  $\Pi : U \rightarrow S_a$  are smooth functions.

Using this result, one can think of characterizing periodic orbits as fixed points of  $\Pi(z) = \phi(\mathcal{T}(z), z)$ . In fact if there exists  $z^* \in S_a$  such that  $\Pi(z^*) = z^*$ , then the flow  $\phi(t, z^*)$  is a periodic orbit with period  $\mathcal{T}(z^*)$ . An example of a Poincaré section is given in Fig. 1.

**Definition 11** Let  $\phi(t, z_0)$  be a periodic solution of (5). The Floquet operator

$$V(t) \doteq \frac{\partial \phi(t, z_0)}{\partial z_0}$$

solves the variational equation

$$\dot{V}(t) = D_z X_H(\phi(t, z_0)) V(t) \quad (15)$$

with  $V(0) = \text{id}$ . The matrix  $V(T)$  is called a monodromy matrix and its eigenvalues are termed Floquet multipliers.

For a general discussion about Floquet theory one can refer to [36,49,59]. In general, the construction of the Poincaré map is not explicit. But knowledge of the monodromy matrix and Proposition 2 with the equation  $\sigma(\mathcal{T}(z), z) = 0$  allow us to compute the Taylor expansion of  $\Pi(z)$  in a neighborhood of  $z_0$  in  $S_a$ . We shall examine this idea in the next section, where we also see the consequences of the fact that periodic orbits in Hamiltonian systems are not isolated. In fact the following result holds true:

**Proposition 3 ([36])** *Periodic solutions are never isolated and  $+1$  is always a multiplier whose eigenvector is  $J\nabla H(z)$ . If  $F: \mathcal{P} \rightarrow \mathbb{R}$  is a first integral, then  $\nabla F$  is a left eigenvector of the monodromy matrix with eigenvalue  $+1$ .*

Note that a consequence is that since in a Hamiltonian system the Hamiltonian  $H$  is always conserved, the multiplier  $+1$  has algebraic multiplicity of at least 2.

In the presence of an integral of motion like the Hamiltonian, the Poincaré map can be restricted to the level set of  $H$ . The map  $\Pi$  turns out to be symplectic [5,36]. The stability of a periodic orbit can now be defined in terms of the stability of the fixed points of the Poincaré map. Let  $\Pi^n(z)$  denote the  $n$ th iteration of the Poincaré map applied to  $z \in S_a$ . We can define

**Definition 12 ([1,36,49])** A periodic orbit  $\phi(\cdot, z_0)$  is stable for all  $\epsilon > 0$  if there exists  $\delta(\epsilon, z_0)$  such that

$$\|z - z_0\| < \delta \text{ implies } \|\Pi^n(z) - z_0\| < \epsilon \quad \text{for all } n > 0.$$

This stability criterion is difficult to verify. A weaker criterion is obtained by considering the linearization of  $\Pi(z)$  at  $z = z_0$ :

**Definition 13 ([1,36])** A periodic orbit  $\phi(\cdot, z_0)$  is spectrally stable if the associated Poincaré map  $\Pi$ , which is restricted to the manifold defined by the integrals of motion, has a linearization  $D_z \Pi(z_0)$  with a spectrum on the unit circle.

By a local change of coordinates one can show that the eigenvectors of  $D_z \Pi(z_0)$  are equal to the eigenvectors of  $V(T_0)$  different from  $X_H(z_0) = J\nabla H(z_0)$  [36]. The Poincaré map removes the degeneracy of the monodromy matrix  $V(T)$ . To illustrate this point let us consider  $x \in \mathbb{R}^n$  and an autonomous system

$$\dot{x}(t) = f(x(t)) \quad (16)$$

with  $f: \mathbb{R}^n \mapsto \mathbb{R}^n$  differentiable. Let  $\phi(t, x_0)$  be a  $T$ -periodic solution of (16) emanating from  $x_0$ . To study the trajectories near  $x_0$  one can construct a local diffeomorphism  $h: \mathbb{R}^n \mapsto \mathbb{R}^n$ ,  $y = h(x)$  such that  $y_0 = h(x_0)$  with (16) in the form

$$\begin{aligned} \dot{y}(t) &= Dh(h^{-1}(y(t))) f(h^{-1}(y(t))) = \tilde{f}(y(t)) \\ \text{with } \tilde{f}(y_0) &= (1, 0, \dots, 0). \end{aligned} \quad (17)$$

In  $y$  coordinates the periodic orbit  $\phi(t, x_0)$  reads  $\psi(t, y_0) = h(\phi(t, h^{-1}(y_0)))$  and we can define a map  $\sigma$  as

$$\sigma(t, y) = \langle \tilde{f}(y_0), \psi(t, y_0) - y \rangle. \quad (18)$$

The form of  $\tilde{f}(y_0)$  implies  $\sigma(y, t) = \psi_1(t, y) - y_{0,1} = h_1(\phi(t, h^{-1}(y_0))) - y_{0,1}$ . An easy calculation shows that

$$\left. \frac{\partial \sigma}{\partial t} \right|_{(0, y_0)} = \sum_{l=1}^n \frac{\partial h_l}{\partial x_l} f_l(x) \Big|_{(0, y_0)} = 1,$$

which implies the existence of a return time  $\tau(y)$  satisfying  $\sigma(\tau(y), y) = 0$  for  $y$  in a sufficiently small neighborhood of  $y_0$ . Now we define the map  $\Pi(y) = \psi(\tau(y), y)$ . By applying the IFT, we can compute

$$\frac{\partial \tau}{\partial y_j} = \delta_{1j} - \sum_{l,m} \frac{\partial h_l}{\partial x_l} \frac{\partial \phi_l}{\partial x_m} \frac{\partial x_m}{\partial y_j} \quad (19)$$

and also the Jacobian of  $\Pi$ :

$$\frac{\partial \Pi_i(y)}{\partial y_j} = \tilde{f}_i(y) \frac{\partial \tau}{\partial y_j} + \sum_{l,m} \frac{\partial y_i}{\partial x_l} \frac{\partial \phi_l}{\partial x_m} \frac{\partial x_m}{\partial y_j}. \quad (20)$$

Using (19), one can easily say that the matrix  $\partial \Pi_i(y)/\partial y_j$  at  $y = y_0$  has the first column equal to  $\tilde{f}(y_0) = (1, 0, \dots, 0)$ . The diffeomorphism  $h$  allowed us to “isolate” the direction of the vector field  $f$  at  $x_0$ . This implies that the map  $\Pi_{S_{\tilde{f}(z_0)}}$ , the restriction of  $\Pi$  on the Poincaré section  $S_{\tilde{f}(z_0)} = \{y: \langle \tilde{f}(y_0), y - y_0 \rangle = 0\} \simeq \mathbb{R}^{n-1}$ , maps  $S_{\tilde{f}(z_0)}$  into  $S_{\tilde{f}(z_0)}$  and its linearization has no eigenvectors in the direction of  $\tilde{f}(y_0)$ . The map  $\Pi_{\tilde{f}(z_0)}$  describes the stability of the periodic orbit  $\phi(t, x_0)$ . A similar construction can be carried out when  $f$  is a Hamiltonian vector field. In that case it is necessary to take into account the existence of integrals of motion and construct the Poincaré map on the manifold where the integral of motions are fixed. A more complete study of the Poincaré map will be presented in the next section.

Let us now present some properties of the Floquet operator. Let us consider a Hamiltonian system with imaginary multipliers. In the linear time-independent case the monodromy matrix is given by

$$V(T_0) = \exp(T_0 JA),$$

where  $T_0 = 2\pi/|\lambda|$ , with  $\lambda$  an imaginary eigenvalue of  $JA$ . Now let  $H = \frac{1}{2}\langle z, Az \rangle + h(z)$  be a Hamiltonian with  $h(z) = o(\|z\|^2)$  near  $z = 0$ . The equation for the monodromy matrix is (15). Let  $\phi_0(t)$  be another periodic orbit, then (15) can be written as follows:

$$\frac{dV_\psi(s)}{ds} = \frac{T_\phi}{T_0} DX_H(\psi_0(s)) V_\psi(s),$$

where  $t = T_0 s / T_\phi$  and  $\psi_0(s) = \phi_0(T_0 s / T_\phi)$ . Then one can show

**Proposition 4** ([39]) *The Floquet operator  $V_\psi(t)$  is  $C^\infty$  in  $\psi$  and  $T_\phi$ .*

**Corollary 1** ([39]) *If  $\phi_0(t)$  is sufficiently close to  $z = 0$ , then  $V_\phi(T_\phi)$  is arbitrarily close to  $\exp(T_0 JA)$ .*

Corollary 1 is interesting because it allows us to use linearized dynamics. Now the analysis of the nonlinear stability can be carried out by using Krein's theory. For this typical references are [21,59]. Here we recall

**Theorem 4 (Krein)** *Let  $V(T)$  be a spectrally stable monodromy matrix. Let  $Q_\lambda$  be a quadratic form  $Q_\lambda(v) = \langle V(T)v, Jv \rangle$  where  $v$  belongs to the  $\lambda$  eigenspace of  $V(T)$ . Then  $V(T)$  is in an open set of spectrally stable matrices if and only if the quadratic  $Q_\lambda$  has definite sign.*

By combining Corollary 1 and Theorem 4, we could show that a solution  $\phi_0(t)$  close enough to  $z = 0$  is spectrally stable.

### Continuation of Periodic Orbits in Hamiltonian Systems

In general it is difficult to prove the existence of and then construct periodic orbits. Sometimes, for certain specific values of the parameters characterizing the system it is possible to find particular solutions. Typically these are the equilibria and relative equilibria. In these cases the continuation method can be a useful approach. The idea is to look at how a given periodic orbit changes according to a small modification of the parameters. The method reduces the research of periodic orbits to the problem of finding fixed points of the Poincaré map that can be continued as a function of the parameters; see Fig. 1.

**Definition 14 (Continuation of an orbit)** Given a dynamical system and  $\phi_0(t)$  one of its orbits, we say that  $\phi_0(t)$  can be continued if there exists a family of orbits  $\phi(t, \alpha)$  smoothly dependent on parameters  $\alpha$ 's and such that  $\phi(t, 0) = \phi_0(t)$ .

Let us consider a Hamiltonian vector field  $X_H(z, \alpha)$  where  $z \in \mathcal{P}$  and  $\alpha \in \mathbb{R}^k$  are  $k$  parameters. We now write the equations of motion in a form where the period  $T$  appears explicitly. After  $t \rightarrow t/T$  the equations read

$$\dot{z}(t) = TJ \nabla H(z(t), \alpha) = TX_H(z(t), \alpha), \quad (21)$$

with  $t \in [0, 1]$ .

**Definition 15** Let  $\phi(t, z, T, \alpha)$  be a solution of (21). The map

$$R(z, T, \alpha) \doteq \phi(1, z, T, \alpha) - z \quad (22)$$

is called a return map.

The orbit  $\phi(t, z_0, \alpha_0)$  is  $T_0$ -periodic if  $R(z_0, T_0, \alpha_0) = 0$ . Now we are interested to see what is the fate of the orbit when  $z_0$ ,  $T_0$  and  $\alpha_0$  are varied; therefore, it is useful to determine the local behavior of the map  $R$ . This is collected in the following proposition.

**Proposition 5** ([42]) *Let  $\phi_0(t, z_0)$  be a periodic orbit with period  $T_0$  and  $\alpha = 0$ , then the following relations hold*

- $D_z R(z_0, T_0, 0) = V(1) - \text{id}$ ,
- $X_H(z_0, 0) \in \ker(V(1) - \text{id})$ ,
- $D_T R(z_0, T_0, 0) = X_H(z_0, 0)$ ,
- $D_\alpha R(z_0, T_0, 0) = T_0 V(T) \int_0^1 V^{-1}(s) D_\alpha X_H(\phi_0(s, z_0)) ds$ ,

together with the differential map

$$\begin{aligned} DR(z_0, T_0, 0)(\xi, T, \alpha) \\ = (V(1) - \text{id}) \cdot \xi + T X_H(z_0, 0) + D_\alpha R(z_0, T_0, 0) \cdot \alpha, \end{aligned}$$

where  $(\xi, T, \alpha) \in \mathbb{R}^{2n+k+1}$ .

Initially let us consider a non-Hamiltonian dynamical system in  $\mathbb{R}^n$ :

$$\dot{x}(t) = f(x(t), \epsilon), \quad \epsilon \in \mathbb{R}. \quad (23)$$

We now illustrate how the IFT is used to construct a new solution from a given one, i.e., by continuation. We present a detailed proof for the reader's convenience.

**Proposition 6** ([3,42]) *Let  $\Gamma_0 = \{\phi_{T_0}(t, x_0), t \geq 0\}$  be a periodic orbit of period 1 of  $\dot{x}(t) = T_0 f(x(t), \epsilon)$  for  $\epsilon = 0$ . If 0 is an eigenvalue of  $D_x R(x_0, T_0, 0)$  with multiplicity 1, then orbit  $\Gamma_0$  can be continued.*

*Proof* Let us consider the map  $G: \mathbb{R}^n \times \mathbb{R}_+ \times [0, 1] \rightarrow \mathbb{R}^{n+1}$  defined by

$$G(x, T, \epsilon) = (R(x, T, \epsilon), \langle f(x_0, 0), x - x_0 \rangle). \quad (24)$$

Note that  $\langle f(x_0), x - x_0 \rangle = 0$  is the equation of the Poincaré section  $S_{f(x_0)}$ . Since  $\Gamma_0$  is a periodic orbit with period  $T_0$  emanating from  $x_0$ , then  $G(x_0, T_0, 0) = 0$ . As already stated the strategy is to employ the IFT to derive the continuation of  $\Gamma_0$ . Thus, it is necessary to compute  $D_{x,T}G$  at  $(x_0, T_0, 0)$ :

$$D_{x,T}G(x_0, T_0, 0) = \begin{pmatrix} D_x R(x_0, T_0, 0) & f(x_0, 0) \\ f(x_0, 0) & 0 \end{pmatrix}.$$

In order to show that  $D_{x,T}G(x_0, T_0, 0)$  is invertible one notes that the equation

$$D_{x,T}G(x_0, T_0, 0)(X, a) = (0, 0)$$

is equivalent to the system

$$\begin{cases} D_x R(x_0, T_0, 0)(X) + f(x_0, 0) a = 0 \\ \langle f(x_0, 0), X \rangle = 0. \end{cases} \quad (25)$$

Assume  $X \neq 0$ . The multiplicity of the 0 eigenvalue of  $D_x R(x_0, T_0, 0)$  is 1 and

$$D_x R(x_0, T_0, 0) f(x_0, 0) = 0.$$

Now from (25) we derive  $(D_x R(x_0, T_0, 0))^2 X = 0$ . Therefore,  $D_x R(x_0, T_0, 0)(X) + f(x_0, 0) a = 0$  would be satisfied only for  $a = 0$  and  $X = cf(x_0, 0)$  with  $c \in \mathbb{R}$ . But this would imply  $c \langle f(x_0, 0), f(x_0, 0) \rangle = 0$ , which is possible only for  $c = 0$ . The kernel of  $D_{x,T}G$  is empty at  $(x_0, T_0, 0)$  and therefore the map is invertible and the application of the IFT provides the existence of  $T(\epsilon)$  and  $x(\epsilon)$  for  $\epsilon$  in a neighborhood of zero such that  $x(0) = x_0$ ,  $T(0) = T_0$  and  $G(x(\epsilon), T(\epsilon), \epsilon) = 0$ . This corresponds to the existence of a new periodic orbit close to  $\Gamma_0$ . Upon the assumption of sufficient regularity for the vector field, the IFT provides also the possibility of approximating  $x(\epsilon)$  and  $T(\epsilon)$  by constructing a Taylor expansion in  $\epsilon$ . Let  $x(\epsilon) = x_0 + \xi(\epsilon)$  and  $T(\epsilon) = T_0 + \tau(\epsilon)$ , then (25) can be evaluated along the continuation curve defined by  $(\xi(\epsilon), \tau(\epsilon), \epsilon)$ :

$$\begin{cases} D_x R(x_0, T_0, 0)(\xi'(\epsilon)) + f(x_0, 0) \tau'(\epsilon) \\ + T_0 V(1) \int_0^1 V(s) \frac{df(\phi_{T_0}(s, x_0), 0)}{d\epsilon} ds = 0 \\ \langle f(x_0, 0), \xi'(\epsilon) \rangle = 0. \end{cases} \quad (26)$$

This set of equations allows us to compute an approximation for  $\xi(\epsilon)$  and  $\tau(\epsilon)$ .  $\square$

For a general dynamical system in  $\mathbb{R}^n$  (23) the possibility of constructing a Poincaré section is related to the notion of a nondegenerate periodic orbit.

**Definition 16** A periodic orbit  $\phi(t, z_0)$  is called nondegenerate if

$$\text{rank}(V(1) - \text{id}) \oplus \mathbb{R} f(z_0) = \mathbb{R}^n.$$

Now for Hamiltonian systems the time evolution is contained in the level set determined by the the Hamiltonian function (the energy) and all integrals of motion  $\{F_i\}_{i=1}^k$

$$X_H(F_i) = \{H, F_i\} = 0, \quad i = 1, \dots, k.$$

This requires a modification of the notion of nondegeneracy. In fact note that the set of integrals of motions  $(F_1, \dots, F_k)$  define a map  $F: \mathcal{P} \mapsto \mathbb{R}^k$  and span  $W = \{\nabla F_i, i = 1 \dots k\}$ . Now we have  $W^\perp = \ker(D_z F(z))$ . One can show  $\text{rank}(V(1) - \text{id}) \subset \ker(D_z F(z))$  and  $X_H(F) = 0$ . The last condition is equivalent to  $X_H(z) \in \ker(D_z F(z))$ . Suppose that  $\dim W^\perp = 2n - k$ , if  $\dim(\ker(V(1) - \text{id})) = k$  then  $\dim(\text{rank}(V(1) - \text{id})) = 2n - k$  and hence  $\text{rank}(V(1) - \text{id}) = W^\perp$ . Therefore, in the Hamiltonian case the natural notion of nondegeneracy has to involve the presence of integrals of motion. This is obtained by using the following definition.

**Definition 17** ([42]) A periodic orbit  $\phi(t, z_0)$  is called normal if

$$\text{rank}(V(1) - \text{id}) \oplus \mathbb{R} X_H(z_0) = W^\perp, \quad (27)$$

where all the gradients  $\{\nabla H, \nabla F_1, \dots, \nabla F_k\}$  are linearly independent.

**Lemma 2** ([42]) If the algebraic multiplicity of the zero eigenvalue of  $V(1) - \text{id}$  is  $m_a = k + 1$ , then condition (27) is satisfied.

Finally we present how to construct the continuation of nontrivial periodic orbits in Hamiltonian systems.

**Theorem 5** ([42]) Let  $\Gamma_0 = \{\phi_{T_0}(t, z_0) : t \in [0, 1]\}$  be an orbit of the Hamiltonian system  $H_0$  on the energy level  $e_0$ . Let the algebraic multiplicity of eigenvalue 0 of  $V(1) - \text{id}$  be 2. Let  $H_\epsilon = H_0 + \epsilon H_1$  be a smooth perturbation of  $H_0$ , then there exists a two-dimensional family of normal periodic orbits  $\phi(t, z_0; \epsilon, e)$  where  $\phi(t, z_0, e_0, 0) = \phi(t, z_0)$ .

*Proof* Let us consider the map  $G: \mathbb{R}^{2n} \times \mathbb{R}_+ \times [0, 1] \rightarrow \mathbb{R}^{2n+1}$  defined by

$$G(x, T, e, \epsilon) = (R(z, T, \epsilon), \langle X_H(z_0, 0), z - z_0 \rangle, H_\epsilon(z) - e). \quad (28)$$

Since  $\Gamma_0$  is a periodic orbit with period  $T_0$  emanating from  $z_0$  then  $G(z_0, T_0, e_0, 0) = 0$ . The strategy is always to employ the IFT to derive the continuation of  $\Gamma_0$ . Thus, it is necessary to compute  $D_{z,T}G$  at  $(z_0, T_0, e_0, 0)$ . A straightforward calculation gives

$$D_{z,T}G(z_0, T_0, 0) = \begin{pmatrix} D_z R(z_0, T_0, 0) & X_H(z_0, 0) \\ X_H(z_0, 0) & 0 \\ \nabla H_0(z_0) & 0 \end{pmatrix}.$$

In order to show that  $D_{z,T}G(z_0, T_0, h_0, 0)$  is invertible one notes that the equation

$$D_{z,T}G(z_0, T_0, h_0, 0)(X, a, 0) = (0, 0, 0)$$

is equivalent to the system

$$\begin{cases} D_z R(z_0, T_0, e_0, 0)(X) + X_H(z_0, 0) a = 0 \\ \langle X_H(z_0, 0), X \rangle = 0 \\ \langle \nabla H_0(z_0), X \rangle = 0. \end{cases} \quad (29)$$

Since  $X_H(z_0) \in \ker D_z R(z_0, T_0, e_0, 0)$  then from (29) we derive  $(D_z R(z_0, T_0, e_0, 0))^2(X) = 0$  and therefore  $a = 0$ . Now Lemma 2 implies that  $X \in W^\perp$ ; therefore, the third equation in (29) is satisfied and the first implies  $X = b X_H(z_0, 0)$  for some  $b \in \mathbb{R}$ . But this would contradict  $b \langle X_H(z_0, 0), X_H(z_0, 0) \rangle = 0$  unless  $b = 0$ . This implies that that kernel of  $D_{x,T}G$  is empty at  $(x_0, T_0, e_0, 0)$  and therefore the map  $G$  is invertible. The application of the IFT provides the existence of  $T(e, \epsilon)$  and  $z(e, \epsilon)$  for  $\epsilon$  in a neighborhood of 0 and  $e$  in a neighborhood of  $e_0$  such that  $z(e_0, 0) = z_0$ ,  $T(e_0, 0) = T_0$ . The functions  $T(e, \epsilon)$  and  $z(e, \epsilon)$  satisfy  $G(z(e, \epsilon), T(e, \epsilon), e, \epsilon) = 0$ . This corresponds to the existence of a new periodic orbit close to  $\Gamma$ . Upon the assumption of sufficient regularity for the vector field, the IFT provides also the possibility of approximating  $z(e, \epsilon)$  and  $T(e, \epsilon)$  by following the same line of argument seen in Proposition 6. This concludes the Proof.  $\square$

**Numerical Studies** The analysis of periodic orbits is very important for its concrete applications, hence for the construction of numerical algorithms to construct periodic orbits and their continuation. Here we do not consider explicitly this problem but the reader is invited to consult, for example, [31,32]. Moreover there is some free and open-source software available, for example, AUTO (see <http://indy.cs.concordia.ca/auto/> and [19,20]).

### Example

On  $\mathcal{P} = \mathbb{R}^4$  with coordinates  $z = (p, q)$  and the standard symplectic form, we consider the Hamiltonian

$$H = \frac{1}{2} \|p\|^2 + V_0(q), \quad V_0(q) = \frac{1}{2} \|p\|^2 - \frac{\lambda}{2} \|q\|^2 + \frac{1}{4} \|q\|^4 \quad (30)$$

with  $\lambda > 0$ . The reader could check that  $V_0(q)$  has the shape of a “Mexican-hat.” The Hamilton equation Hamiltonians of motion read

$$\begin{cases} \dot{q} = p \\ \dot{p} = (\lambda - \|q\|^2)q. \end{cases} \quad (31)$$

Let  $e(t)$  be a unit vector with  $e(t + 2\pi/\nu_0) = e(t)$ , then there is a periodic orbit of the form  $q_0(t) = A_0 e(t)$ ,

$p_0(t) = A_0 \dot{e}(t)$ . This orbit is a relative equilibrium (see below for a formal definition). The initial conditions determine  $A_0, \nu_0$  and in particular the energy  $E$ , which in turn can be used to parameterize  $A_0, \nu_0$ :  $A_0^2 = \frac{2}{3}(\lambda + \sqrt{\lambda^2 + 3E})$ ,  $\nu_0^2 = \frac{1}{3}(-\lambda + \sqrt{\lambda^2 + 3E})$ ; here  $E \leq 0$ . Note that an easy calculation shows that the admissible values of the energy are  $E \geq \min_q \{V_0(q)\} = -\lambda^2/4$ . Let  $\phi(t, z_0) = (p_0(t), q_0(t))$  be the periodic solution, then the Floquet operator is

$$V(t) = \frac{\partial \phi(t, z_0)}{\partial z_0}$$

and the linearized equations can be written as

$$\frac{dV(t)}{dt} = M(t)V(t), \quad \text{where } M(t) \text{ is a } 4 \times 4 \text{ matrix:}$$

$$M(t) = \begin{pmatrix} 0 & \text{id} \\ -D^2 V_0(q_0(t)) & 0 \end{pmatrix}.$$

Now if we perform the transformation

$$W(t) = S(t) V(t) = \begin{pmatrix} R(t) & 0 \\ 0 & R(t) \end{pmatrix} X(t)$$

$$\text{where } R(t) = \begin{pmatrix} \cos \nu_0 t & -\sin \nu_0 t \\ \sin \nu_0 t & \cos \nu_0 t \end{pmatrix}$$

we obtain

$$\frac{dW(t)}{dt} = \hat{M}W(t),$$

where

$$\hat{M} = \dot{S}(t) S^{-1}(t) + S(t) M(t) S^{-1}(t).$$

$\hat{M}$  is not time-dependent. In fact  $\hat{M}$  is the following matrix:

$$\hat{M} = \begin{pmatrix} \Omega & \text{id} \\ -R^T(t) D^2 V_0(q_0(t)) R(t) & \Omega \end{pmatrix},$$

where

$$\Omega = \begin{pmatrix} 0 & -\nu_0 \\ \nu_0 & 0 \end{pmatrix} \quad \text{and}$$

$$-R^T(t) D^2 V_0(q_0(t)) R(t) = \begin{pmatrix} -3A_0^2 + \lambda & 0 \\ 0 & -A_0^2 + \lambda \end{pmatrix}.$$

Now at  $t = T_0$

$$V(T_0) = S^{-1}(T_0) W(T_0) = \exp(\hat{M} T_0).$$

The matrix  $\hat{M}$  has spectrum

$$\text{Spec}(\hat{M}) = \left\{ 0, 0, \pm i \sqrt{6A_0^2 - 4\lambda} \right\}$$



and the corresponding multipliers are:

$$\begin{aligned}\exp(\text{Spec}(\hat{M})) &= \left\{ 1, 1, \exp\left(\pm i T_0 \sqrt{6A_0^2 - 4\lambda}\right) \right\} \\ &= \left\{ 1, 1, \exp\left(\pm i T_0 \sqrt[4]{\lambda^2 + 3E}\right) \right\},\end{aligned}$$

where  $T_0 = 2\pi/\nu_0$ . The matrix  $V(T_0) - \mathbf{id}$  has a zero eigenvalue with multiplicity 2, and the periodic orbit  $\phi_0(t, z_0)$  can be continued by using Theorem 5.

**Remark 5** Note that using the expression for  $\nu_0$ , one can check that for  $E = E_n$ , where

$$E_n = \frac{\lambda^2}{3} \left[ -1 + \frac{n^4}{(n^2 - 12)^2} \right] \text{ with } n \in \mathbb{Z},$$

the third and the fourth multipliers coalesce to 1 and therefore  $q_0(\cdot)$  loses its linear stability in the directions transverse to itself.

### Hamiltonian Systems with Symmetries

In many applications there are Hamiltonian systems with symmetries, the best known example of which is the  $N$ -body problem (see [1,5] and ► [n-Body Problem and Choreographies](#)). A simpler example is (31), which is symmetric with respect to the linear action of  $SO(2)$ . For such systems the Hamiltonian function is invariant under the action of a Lie group  $G$ . We shall see that symmetries can greatly simplify the study of the dynamics. A Hamiltonian system with symmetry is a quadruple  $(\mathcal{P}, \omega, H, G)$  [1,5,48], where

- $(\mathcal{P}, \omega)$  is a symplectic manifold,
- $H: \mathcal{P} \rightarrow \mathbb{R}$  is a Hamiltonian function,
- $G$  is a Lie group that acts smoothly on  $\mathcal{P}$  according to  $G \times \mathcal{P} \ni (g, z) \mapsto \Phi_g(z) \in \mathcal{P}$ . The map  $\Phi_g$  preserves the Hamiltonian ( $G$ -invariance) that is  $H(\Phi_g(z)) = H(z)$ .
- The action  $\Phi$  is semisymplectic, namely,  $\Phi_g^* \omega = \chi(g)\omega$ , with  $\chi(g) = \pm 1$ .  $\chi(\cdot)$  is called temporal character.

In the sequel we consider symplectic actions whereby  $\chi(g) = 1$  for all  $g \in G$ . One can show that for a system  $(\mathcal{P}, \omega, H, G)$  the vector field  $X_H(z)$  is *equivariant* [1], that is,

$$X_H(\Phi_g(z)) = D_z \Phi_g(z) X_H(z). \quad (32)$$

With any element  $\xi$  in the Lie algebra  $\mathfrak{g}$  of  $G$  we can associate a infinitesimal generator  $\xi_{\mathcal{P}}(z)$  of the action defined by

$$\xi_{\mathcal{P}}(z) = \left. \frac{d\Phi_{\exp(\xi t)}(z)}{dt} \right|_{t=0}. \quad (33)$$

**Remark 6** In many applications Hamiltonian systems are constructed from Lagrangian systems; therefore, a symmetry appears usually as a group action on the configuration space  $\mathcal{M}$ . The Hamiltonian symmetry is then the *lifted* action to the cotangent bundle  $T^*\mathcal{M}$ . For instance, if  $SO(2)$  acts linearly on  $\mathcal{M} = \mathbb{R}^2$  by  $\Psi_R(q) = Rq$  with  $R \in SO(2)$ , then its lifted action on  $T^*\mathcal{M} \simeq \mathbb{R}^4$  is  $\Phi_R(q, p) = (Rq, R^T p)$ , where  $R^T$  is the transpose of  $R$ . For more details see [1,36,48].

### Symmetry and Reduction

Given a symplectic action of a group  $G$  there is a map  $\mathbf{J}: \mathcal{P} \rightarrow \mathfrak{g}^*$  defined by

$$\begin{aligned}\langle \mathbf{J}(v(z)), \xi \rangle &= \omega(v(z), \xi_{\mathcal{P}}(z)) \\ \text{for all } z \in \mathcal{P}, v \in T_z \mathcal{P} \text{ and } \xi \in \mathfrak{g}.\end{aligned} \quad (34)$$

The map  $\mathbf{J}$  is called a *momentum map*. This always exists locally, and its global existence requires conditions on  $G$  and the topology of  $\mathcal{P}$  [27]. The crucial property of the momentum map is its encoding of the conserved quantities associated with the  $G$  action. This is the content of the famous Noether's theorem that reads in modern formulation as follows.

**Theorem 6 (Noether [1,5])** *Let  $H$  be a  $G$ -invariant Hamiltonian on  $\mathcal{P}$  with momentum map  $\mathbf{J}$ . Then  $\mathbf{J}$  is conserved on the trajectories of the Hamiltonian vector field  $X_H$ .*

For instance, in a system like (30) with a  $SO(2)$  symmetry action the momentum map is the classical angular momentum

$$\mathbf{J}(p, q) = p \wedge q.$$

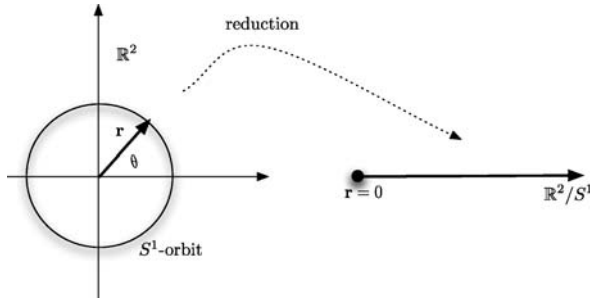
The momentum map  $\mathbf{J}$  has also the property of being equivariant with respect to the coadjoint action associated with  $G$ . In the case of  $G$  being semisimple or compact the result is as follows.

**Theorem 7 (Souriau [27])**

$$\mathbf{J}(\Phi_g(z)) = \text{Ad}_g^* \mathbf{J}(z).$$

The level sets of the momentum map are invariant with respect to the Hamiltonian flow; thus, it is natural to restrict the motion to  $\mathbf{J}^{-1}(\mu)$ . The construction of the dynamics reduced to the manifold defined by the conserved quantities is the origin of the theory of *symmetry reduction*. The first important result is the following.

**Theorem 8 (Marsden–Weinstein)** *Let  $(\mathcal{P}, \omega, H, G)$  be a Hamiltonian system with a symplectic action of the Lie*



Periodic Orbits of Hamiltonian Systems, Figure 2

A simple example: the reduction of  $S^1$  group action on the plane is singular. There are two strata  $r = 0$  and  $r \in (0, \infty)$

group  $G$ , then the triple  $(\mathcal{P}_\mu, \omega_\mu, H_\mu)$  is called a reduced Hamiltonian system where

$$\pi: \mathcal{P} \rightarrow \mathcal{P}_\mu = \mathbf{J}^{-1}(\mu)/G_\mu, \quad \pi^*\omega = \omega_\mu, \quad H_\mu = H \circ \pi$$

and

$$G_\mu = \{g \in G: \text{Ad}_g^*(\mu) = \mu\}.$$

The manifold  $\mathcal{P}_\mu$  is symplectic with symplectic form  $\omega_\mu$ . The Hamiltonian flow on  $\mathcal{P}_\mu$  is induced by the Hamiltonian vector field  $X_{H_\mu}$  defined by

$$i_{X_{H_\mu}} \omega_\mu = dH_\mu.$$

**Remark 7** Note that the reduced vector field  $X_{H_\mu}$  is now defined on a manifold whose dimension is  $\dim \mathcal{P}_\mu = \dim \mathcal{P} - \dim G - \dim G_\mu$ .

For a detailed discussion of this theorem the reader should refer to [1,16,38]. In many cases it can be more useful to perform the reduction through a dual approach using the description of the dynamics in terms of functions on  $\mathcal{P}$ , namely, through the Poisson approach. This permits us to consider also cases where  $\mathcal{P}_\mu$  is not a smooth manifold but rather a stratified space. A very simple example is shown in Fig. 2.

The theory of singular reduction can be found in [4]. A recent exposition and generalization is given in [45]. The reader could, in particular, consider the expositions given in [16,17], where it is shown through several examples that invariant theory and algebraic methods can be applied to describe reduced dynamics on spaces with singularities. The singular reduction can be summarized in the following result:

**Theorem 9 (Singular reduction [45])** Let  $(\mathcal{P}, \{.,.\})$  be a Poisson manifold and let  $\Phi: G \times \mathcal{P} \rightarrow \mathcal{P}$  be a smooth proper action preserving the Poisson bracket. Then the following holds:

- (i) The pair  $(\mathcal{F}(\mathcal{P}/G), \{.,.\}_{\mathcal{P}/G})$  is a Poisson algebra, where the Poisson bracket  $\{.,.\}_{\mathcal{P}/G}$  is characterized by  $\{f, g\}_{\mathcal{P}/G} = \{f \circ \pi, g \circ \pi\}$ ; for any  $f, g \in \mathcal{F}(\mathcal{P}/G)$ ;  $\pi: \mathcal{P} \rightarrow \mathcal{P}/G$  denotes the canonical smooth projection.
- (ii) Let  $h$  be a  $G$ -invariant function on  $M$ . The Hamiltonian flow  $\phi(t, .)$  of  $X_h$  commutes with the  $G$ -action, so it induces a flow  $\phi^{P/G}(.,.)$  on  $\mathcal{P}/G$  which is a Poisson flow and is characterized by  $\phi(t, z) = \phi^{P/G}(t, \pi(z))$ .
- (iii) The flow  $\phi^{P/G}(.,.)$  is the unique Hamiltonian flow defined by the function  $[h] \in \mathcal{F}(\mathcal{P}/G)$  defined by  $[h](\pi(z)) = h(z)$ . We will call  $H_{P/G} = [h]$  the reduced Hamiltonian.

### Relative Equilibria, Relative Periodic Orbits and their Continuation

In a Hamiltonian system with symmetry there are orbits that originated just from the symmetry invariance. These are the relative equilibria.

**Definition 18 (Relative equilibria)** A curve  $z_e(t)$  in  $\mathcal{P}$  is a relative equilibrium of  $(\mathcal{P}, \omega, H, G)$  if  $z_e(t) = \Phi_{g(t)}(w_e)$ , where  $g(t)$  is a curve in  $G$  and  $w_e$  is such that  $X_H(z_e) = \dot{z}_e(t)$ .

Note that if  $g(t) = g(t + T)$  then  $z_e(t)$  becomes a  $T$ -periodic orbit. For instance in the system (30) there is  $z_e(t) = R(t)w_e = (A_0 e(t), A_0 \dot{e}(t))$  with  $R \in SO(2)$ . There are orbits that can be considered closed up to a  $G$ -action, i. e., they are closed on  $\mathcal{P}/G_\mu$ . These are the relative periodic orbits.

**Definition 19 (Relative periodic orbit)** A curve  $z(t)$  in  $\mathcal{P}$  is a relative periodic orbit of  $(\mathcal{P}, \omega, H, G)$  with period  $T$  if  $z(t + T) = \Phi_g(z(t))$ , where  $g \in G$  and  $g \neq \text{id}$ .

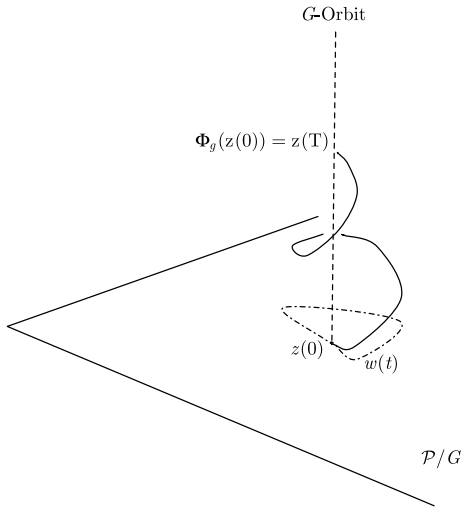
An illustration is given in Fig. 3.

The reduction theory allows us to redefine the relative equilibrium. Let  $\mu \in \mathfrak{g}$  and  $z \in \mathcal{P}_\mu$  be an equilibrium for  $X_{H_\mu}$ . Then  $z$  gives rise to a relative equilibrium in  $\mathcal{P}$ , for a closed curve  $g(t)$  in  $G_\mu$  and define  $w(t) = \Phi_{g(t)}(w_e)$  with  $w_e \in \pi_\mu^{-1}(z)$ . If  $z(t)$  is a relative equilibrium then  $\pi(z(t))$  is an equilibrium of the reduced flow [1]. In general there are two possible ways to study the relative equilibrium. The first approach is to construct the reduction of the Hamiltonian system and then to study

$$X_{H_\mu}(z) = 0 \quad (35)$$

in Hamiltonian form and

$$\{H_G, f\}_{P/G}(z) = 0 \quad \text{for all } f \in \mathcal{F}(\mathcal{P}/G) \quad (36)$$



**Periodic Orbits of Hamiltonian Systems, Figure 3**

A relative periodic orbit. After a relative period  $T$  the orbit  $z(\cdot)$  closes onto the group orbit. The projection of  $z(\cdot)$  on  $\mathcal{P}/G$  is closed

in the Poisson description. The second approach is to observe that if  $z(t) = \Phi_{g(t)}(z^*)$  is a relative equilibrium, then the condition  $\dot{z}(t) = X_H(z(t))$  implies

$$\begin{aligned} dH(z^*) - d\langle J(z^*), \xi \rangle &= 0, \\ \text{with; } \xi_P(z^*) &= \left. \frac{d\Phi_{g(t)}(z^*)}{dt} \right|_{t=0}. \end{aligned} \quad (37)$$

On the three formulations, (35), (36) and (37), one can apply all the standard continuation techniques based on the IFT. The possible difficulty in the study of the continuation of relative equilibria and relative periodic orbits can be seen by looking at the linearization of (37). The symmetry contributes to the degeneracy of the linearized equations, in fact one can show

**Proposition 7 ([37])** *Let  $X(z)$  be an equivariant vector field with respect to a Lie group  $G$ . Then a relative equilibrium has multiplier  $+1$  with multiplicity at least equal to  $\dim \mathfrak{g}$ ; and a relative periodic orbit has multiplier  $+1$  with multiplicity at least equal to  $\dim \mathfrak{g} + 1$ .*

In [39] relative periodic orbits are defined by looking at the fixed points of the action of  $G \times S^1$  on the space of  $T$ -periodic paths. This reads

$$(g, \theta) \cdot z(t) = \Phi_g(z(t + \theta T)). \quad (38)$$

**Remark 8** Note that (38) depends on  $T$  and it is easy to verify that  $T/k$  is a minimal period if the intersection of the isotropy subgroup of the action  $G \times S^1$  with  $S^1$  is equal to  $\mathbb{Z}_k$ .

**Theorem 10 ([39])** *Let  $(\mathcal{P}, \omega, H, G)$  be a symmetric Hamiltonian system. Let  $z_e \in \mathcal{P}$  such that*

1.  $d^2 H(z_e)$  is a nondegenerate quadratic form
2.  $d^2 H(z_e)$  is positive-definite if restricted to  $V_\lambda$ , which is the real part of the eigenspace associated with the eigenvalue  $i\lambda$ ,  $\lambda \in \mathbb{R}$ .

*Then for every isotropy subgroup  $\Gamma$  of the  $G \times S^1$ -action on  $V_\lambda$  and  $\epsilon$  sufficiently small there exist at least  $\text{Fix}(\Gamma, V_\lambda)$  periodic trajectories with period near  $2\pi/|\lambda|$ , a symmetry group contained in  $\Gamma$  and  $|H(z(t)) - H(z_e)| = \epsilon^2$ .*

This result has been proved using a combination of a variational approach and the IFT, and will be considered in “Hamiltonian Viewpoint”. We have seen that the Poincaré section is a useful construction to analyze the local structure of a periodic orbit. A very similar approach can be introduced for relative periodic orbits. A relative periodic orbit can be seen as a combination of motions in  $\mathcal{P}$  and in the symmetry group  $G$ . It is therefore natural to look for a suitable decomposition of the motion. In [39] following decomposition was introduced:

$$T_z \mathcal{P} = W \oplus X \oplus Y \oplus Z, \quad (39)$$

where

$$\begin{aligned} W &= \ker D_z J(z) \cap T_z(G \cdot z), \quad G \cdot z = \{\Phi_g(z) : g \in G\}, \\ X &= T_z(G \cdot z)/W, \\ Y &= \ker D_z J(z)/W, \\ Z &= T_z \mathcal{P}/(\ker D_z J(z) + T_z(G \cdot z)). \end{aligned}$$

Let  $G_z = \{g \in G : \Phi_g(z) = z\}$  and  $G_\mu = \{g \in G : \text{Ad}_g^*(\mu) = \mu\}$  be the isotropy subgroups. It turns out that

- $\ker D_z J(z)$  and  $T_z(G \cdot z)$  are  $\omega$ -orthogonal,
- $\omega$  restricted  $\ker D_z J(z) + T_z(G \cdot z)$  is singular and  $W$  is the kernel,
- $\omega$  induces a  $G_z$ -invariant symplectic form  $\omega_X$  on  $X$  and  $\omega_Y$  on  $Y$ ,
- $\omega$  defines a  $G_z$ -invariant isomorphism between  $W$  and  $Z^*$  the dual of  $Z$ .

The splitting (39) allows us to decompose the vector field  $X_H$  and to analyze the motion along the group orbit and along the transverse directions. This is a tool used in [39] to study the Floquet operator in a neighborhood of the relative periodic orbit. For applications in the study of nonlinear normal modes and stability see [40,41]. In [58] it is shown that such a construction can be used to decompose the Poincaré section in a part which is tangent to the conserved momentum, another part which is tangent to

shape and a part parameterizing the momentum. This approach has also been applied to the study of the geometry of mechanical systems defined on the cotangent bundle of a differentiable manifold; for this see [50] and references therein.

### The Variational Principles and Periodic Orbits

The idea behind variational principles is to transform a problem in differential equations into a question about critical points of a certain functional called *action*, whose domain is formed by trajectories of interest. Now in the study of  $T$ -periodic orbits we are interested in finding trajectories  $\phi$  solving the equations of motion and satisfying  $\phi(t) = \phi(t + T)$ . This is a periodicity condition. The variational approach is particularly useful in the study of periodic problems because the periodicity condition is included in the definition of the space where the action is defined. Furthermore the method enables us to prove results not restricted to *small perturbations*. Let  $\mathcal{A}: \Lambda \mapsto \mathbb{R}$  be a functional on a space of loops usually modeled on a Banach or Hilbert space. We shall see that the condition of vanishing of the first derivative of  $\mathcal{A}$  is equivalent, in an appropriate sense, to solving the equations of motion. In order to describe the properties of  $\mathcal{A}$  let us introduce

**Definition 20 (Critical points)** Let  $\mathcal{A}[\cdot]$  be a differentiable functional on  $\Lambda$ . A path  $q(\cdot)$  is a *critical point* of  $\mathcal{A}[\cdot]$  if

$$D\mathcal{A}[q](v) = 0 \quad \text{for all } v(\cdot) \in T\Lambda.$$

**Definition 21 (Critical set)** The set  $K = \{q(\cdot) \in \Lambda: D\mathcal{A}[q] = 0\}$  is the *critical set* of  $\mathcal{A}[\cdot]$ .

**Definition 22 (Critical value)** A real number  $c$  is called a *critical value* if  $K_c \doteq \mathcal{A}^{-1}[c] \cap K \neq \emptyset$ .

### Lagrangian Viewpoint

A typical way to write Newton's equations in a variational form is by using Lagrange's equations which are formulated through the *least action principle*. This can be achieved for all mechanical systems that have potential forces. Consider a mechanical system whose configuration space is a Riemannian manifold  $\mathcal{M}$  of dimension  $n$ . We denote with  $(q, v_q)$  the local coordinates in  $T\mathcal{M}$  and with  $L: T\mathcal{M} \rightarrow \mathbb{R}$  the Lagrangian function. In particular in the case of the so-called *natural mechanical system* [1] the Lagrangian has the form

$$L(q, v_q) = \frac{1}{2} \langle v_q, v_q \rangle - V(q), \quad (40)$$

where  $\langle \cdot, \cdot \rangle$  is the metric on  $T\mathcal{M}$  and  $V: \mathcal{M} \rightarrow \mathbb{R}$  is the potential. Given a path  $q: [0, T] \rightarrow \mathcal{M}$  with integrable time derivative one can define

**Definition 23 (Action functional in Lagrangian form)**

$$\mathcal{A}_L[q] = \int_0^T dt L(q(t), \dot{q}(t)). \quad (41)$$

In what follows we will be mostly interested in closed paths. Let  $C^2([0, T], \mathbb{R}^n)$  be the space of a closed path of period  $T$  with two continuous time derivatives. One can easily show that

**Proposition 8 ([1,5])** Let  $L$  be a smooth Lagrangian function on  $\mathcal{M}$  of the form (40). Let  $\mathcal{A}_L$  be defined over  $\Lambda = C^2([0, T], \mathcal{M})$ . Then

$$D\mathcal{A}_L[q](v) = 0 \quad \text{for every } v(\cdot) \in T\Lambda \simeq C^2([0, T], T\mathcal{M})$$

is equivalent to Newton's equations with  $q(0) = q(T)$ . In particular the equations of motion in local coordinates are the Euler–Lagrange equations

$$\frac{d}{dt} \frac{\partial L(\dot{q}(t), q(t))}{\partial \dot{q}_i} - \frac{\partial L(\dot{q}(t), q(t))}{\partial q_i} = 0, \quad i = 1, \dots, n.$$

**Remark 9** For historical reasons the condition  $D\mathcal{A}_L[q](v) = 0$  is called *least action* although it is only a condition for stationary points of  $\mathcal{A}_L[\cdot]$  in  $\Lambda$ .

In general the Lagrangian contains a quadratic form in  $v_q$ ; in turn the action has a time integral of a quadratic form in  $\dot{q}(t)$ . This essentially shows that the natural domain for  $\mathcal{A}_L[\cdot]$  is the Sobolev space  $H^1([0, T], \mathbb{R}^n)$ . Actually the action is defined on  $H^1([0, T], \mathcal{M})$ , which is a Hilbert manifold [8,29]. This is locally described by absolutely continuous functions with the time derivative in the Lebesgue space  $L^2$  [28]. The interest in variational methods is related to the possibility of using critical point theory to find critical points corresponding to certain type of trajectories and then to show that such trajectories are solutions of Newton's equations [2,21,33]. In particular this approach has been very successful in studying the problem of periodic orbits. Here is a general result in the Lagrangian setting:

**Theorem 11 ([8])** Let  $\mathcal{M}$  be compact Riemannian manifold and let  $L: T\mathcal{M} \rightarrow \mathbb{R}$  be a Lagrangian of the form (40) with  $V \in C^1(\mathcal{M}, \mathbb{R})$ . Then there exists a periodic orbit in any free homotopy class.

We illustrate the ideas of the proof by considering the special case where  $\mathcal{M}$  is the  $n$ -dimensional torus  $T^n \simeq \mathbb{R}^n/\mathbb{Z}^n$ . The problem is now to show that the action  $\mathcal{A}_L[\cdot]$  attains a critical point, a minimum, in  $\Lambda(\mathcal{M}) =$

$\{q(\cdot) \in H^1([0, T], \mathcal{M}): q(\cdot) \text{ is not null homotopic}\}$ . Let us consider the sublevel of the action  $A_k = \{q(\cdot) \in \Lambda(\mathcal{M}): \mathcal{A}_L[q] \leq k\}$ . Now since  $V$  is smooth there exists  $\min_{\mathcal{M}} V(q) = m > -\infty$  and therefore

$$\mathcal{A}_L[q] \geq \frac{1}{2} \int_0^T dt \|\dot{q}(t)\|^2 - m.$$

Since  $\mathcal{M}$  is compact, the norm of  $q(\cdot)$  in  $\Lambda(\mathcal{M})$  is equivalent to  $\|\dot{q}\|_2$ . This allows us to show that the action  $\mathcal{A}[\cdot]$  is *coercive*, namely:

**Definition 24 (Coercivity)**  $\mathcal{A}: \Lambda \rightarrow \mathbb{R}$  is coercive in  $\Lambda$  if for all  $q_n(\cdot)$  such that  $\lim_{n \rightarrow \infty} \|q_n\|_{\Lambda} = \infty$ . Then  $\lim_{n \rightarrow \infty} \mathcal{A}[q_n] = 0$ .

Moreover in  $A_k$  necessarily we have

$$\|\dot{q}\|_2^2 = \int_0^T dt \|\dot{q}(t)\|^2 \leq 2(k + m).$$

This condition guarantees that  $A_k$  is weakly compact in the topology of  $\Lambda$  [2,33,53] and using the coercivity, one can obtain the existence of a minimizer  $q_*(\cdot)$ . The minimizer  $q_*(\cdot)$  is attained in  $A_k$  and it is a *weak* solution of the equations of motion. Indeed  $\mathcal{A}[q^*] = \min_{\Lambda} \mathcal{A}[q]$  and

$$D\mathcal{A}[q_*](v) = 0 \quad \forall v(\cdot) \in \Lambda(T\mathbb{R}^n),$$

namely,

$$\int_0^T dt \sum_{i=1}^n \left( \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i} v_i(t) + \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial q_i} \dot{v}_i(t) \right) = 0. \quad (42)$$

Using the fact that  $q_*(\cdot)$  is absolutely continuous and that  $L$  is regular, one can integrate by parts

$$\int_0^T dt \sum_{i=1}^n \left( \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i} - \int_0^t ds \frac{\partial L(\dot{q}_*(s), q_*(s))}{\partial q_i} \right) \dot{v}_i(t) = 0, \quad (43)$$

and obtain

$$\frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i} - \int_0^t ds \frac{\partial L(\dot{q}_*(s), q_*(s))}{\partial q_i} = c_i,$$

where  $c_i$  is constant in  $L^2([0, T], \mathbb{R}^n)$ .

By differentiating with respect to  $t$ , we obtain the Euler-Lagrange equations:

$$\frac{d}{dt} \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i} - \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial q_i} = 0 \quad \text{a.e. } \forall i. \quad (44)$$

It turns out that the equality in (44) holds for all times  $t$  because  $L$  is smooth. The reader should observe that the derivations of (42) and (44) are part of the proof of Proposition 8. It is necessary to note that constant paths are not in  $\Lambda(\mathcal{M})$ ; indeed the number of minimizers is bounded from below by the Lusternik-Schnirelman category  $\text{Cat}(\mathcal{M}) = n + 1$  [15,33]. Thus, we exclude possible minimizers which would be trivial periodic orbits. In this very simple example we can therefore appreciate the role of the definition of the space of paths  $\Lambda(\mathcal{M})$  and its topology. More interesting cases can be found in the study of the  $N$ -body problem and in particular in the article [► \*n\*-Body Problem and Choreographies](#) in this encyclopedia.

The result in [8] can be generalized to Lagrangian systems with symmetries. Assume there is a group action on the configuration space  $G \times \mathcal{M} \mapsto \mathcal{M}$  – denoted by  $(g, m) \mapsto g.m$  – which preserves the Lagrangian  $L: T\mathcal{M} \rightarrow \mathbb{R}$  and consider the problem of finding relative periodic paths  $\gamma(\cdot)$  as critical points of the action  $\mathcal{A}_L[\cdot]$ . We need to study the topology of

$$\Lambda^g(\mathcal{M}) = \{\gamma(\cdot) \in H^1([0, T], \mathcal{M}): \gamma(t + T) = g.\gamma(t)\}. \quad (45)$$

In fact if the action  $\mathcal{A}_L[\cdot]$  is bounded from below on  $\Lambda^g(\mathcal{M})$ , then each connected component would contain at least a minimum that is a critical point and therefore a periodic orbit. The following analysis was presented in [35]. In what follows for notational simplicity we denote a path and its homotopy class by the same symbol and use  $*$  to denote both concatenation of paths and the induced operations on homotopy classes. Assume that  $\mathcal{M}$  is connected. Choose a base point  $m \in \mathcal{M}$  and let

$$\Lambda_m^g(\mathcal{M}) \doteq \{\gamma \in \Lambda^g(\mathcal{M}): \gamma(0) = m\},$$

the space of continuous paths from  $m$  to  $gm$ . Let  $\Lambda_m(\mathcal{M}) \doteq \Lambda_m^{\text{id}}(\mathcal{M})$  denote the space of continuous loops based at  $m$ . Note that the space of connected components of  $\Lambda_m(\mathcal{M})$  is the fundamental group of  $\mathcal{M}$ :  $\pi_0(\Lambda_m(\mathcal{M})) = \pi_1(\mathcal{M}, m)$ . Fix a particular path  $\omega \in \Lambda_m^g(\mathcal{M})$ . The map  $\Phi_\omega(\gamma) = \omega^{-1} * \gamma$  is a bijection

$$\Phi_\omega: \pi_0(\Lambda_m^g(\mathcal{M})) \rightarrow \pi_0(\Lambda_m(\mathcal{M})) = \pi_1(\mathcal{M}, m),$$

where  $\omega^{-1}$  is the path obtained by traversing  $\omega$  “backwards.” This bijection depends (only) on the homotopy class of  $\omega$  in  $\Lambda_m^g(\mathcal{M})$ . For any  $\alpha \in \Lambda_m(\mathcal{M})$  let  $g.\alpha$  be the loop in  $\Lambda_{gm}(\mathcal{M})$  obtained by applying the diffeomorphism  $g$  to  $\alpha$  and define an automorphism of  $\pi_1(\mathcal{M}, m)$  by

$$\alpha \mapsto \alpha_g = \omega^{-1} * g.\alpha * \omega.$$



Again this depends (only) on the homotopy class of  $\omega$  in  $\Lambda_m^g(\mathcal{M})$ . Now define the  $g$ -twisted action of  $\pi_1(\mathcal{M}, m)$  on itself by

$$\alpha \cdot \beta = \alpha_g * \beta * \alpha^{-1} \quad \alpha, \beta \in \pi_1(\mathcal{M}, m). \quad (46)$$

The number of connected components of the relative loop space is given by the following result:

**Theorem 12** *The map  $\Phi_\omega$  induces a bijection*

$$\pi_0(\Lambda^g(\mathcal{M})) \cong \overline{\pi_1(\mathcal{M}, m)^g},$$

where  $\overline{\pi_1(\mathcal{M}, m)^g}$  is the set of orbits of the  $g$ -twisted action of  $\pi_1(\mathcal{M}, m)$  on itself.

**Remark 10** If  $g$  is homotopic to the identity then  $\Lambda^g(\mathcal{M})$  is homotopy-equivalent to the loop space  $\Lambda(\mathcal{M}) \doteq \Lambda^{\text{id}}(\mathcal{M})$  and the  $g$ -twisted action of  $\pi_1(\mathcal{M}, m)$  on itself is just conjugation. This is therefore the well-known result that the connected components of the loop space map bijectively to the conjugacy classes of the fundamental group [29].

**Remark 11** The  $g$ -twisted action of  $\pi_1(\mathcal{M}, m)$  on itself induces an affine action of the first homology group  $H_1(\mathcal{M})$  on itself:

$$\langle \alpha \rangle \cdot \langle \beta \rangle = g \cdot \langle \alpha \rangle - \langle \alpha \rangle + \langle \beta \rangle,$$

where the brackets  $\langle \cdot \rangle$  denote the homology class and  $g \cdot \langle \alpha \rangle$  denotes the natural action of  $g$  on  $H_1(\mathcal{M})$ . When  $\pi_1(\mathcal{M}, m)$  is Abelian this is the same as the action of  $\pi_1(\mathcal{M}, m)$  on itself. More generally it is easier to calculate than the  $\pi_1(\mathcal{M}, m)$  action and in typical applications may be used to describe relative periodic orbits in terms of winding numbers.

The analysis gives explicit results for systems whose configuration space has the property that all its homotopy groups except the fundamental group are trivial. In this case  $\mathcal{M}$  is said to be  $K(\pi, 1)$ ; for more details see [12,57]. Examples of  $K(\pi, 1)$  spaces include tori, the plane  $\mathbb{R}^2$  with  $N$  points removed, and the configuration spaces of planar  $N$ -body problems.

**Theorem 13** *Assume  $\mathcal{M}$  is a  $K(\pi, 1)$ . Then for any  $\gamma \in \Lambda_m^g(\mathcal{M})$  the connected component of  $\Lambda^g(\mathcal{M})$  containing  $\gamma$ , denoted  $\Lambda_\gamma^g(\mathcal{M})$ , is also a  $K(\pi, 1)$  with*

$$\pi_1(\Lambda_\gamma^g(\mathcal{M})) \cong Z_{\pi_1(\mathcal{M})}^g(\Phi_\omega(\gamma)),$$

where

$$\begin{aligned} Z_{\pi_1(\mathcal{M})}^g(\Phi_\omega(\gamma)) \\ \doteq \{ \alpha \in \pi_1(\mathcal{M}) : \alpha_g * \Phi_\omega(\gamma) * \alpha^{-1} = \Phi_\omega(\gamma) \} \end{aligned}$$

i.e., the isotropy subgroup (or centralizer) at  $\Phi_\omega(\gamma)$  of the  $g$ -twisted action of  $\pi_1(\mathcal{M}, m)$  on itself.

We note that all  $K(\pi, 1)$ 's with isomorphic fundamental groups are homotopy-equivalent to each other [12,57], and so this result determines the homotopy types of connected components of relative loop spaces. The homology groups can be computed algebraically as the homology groups of the fundamental group [14].

**A Simple Example** Let  $\mathcal{M} = T^1$ , the circle, and consider first the loop space  $\Lambda(T^1)$ . The “ $g$ -twisted action” of  $\pi_1(T^1)$  on itself is just conjugation, and since  $\pi_1(T^1) \cong \mathbb{Z}$  is Abelian this is trivial. So  $\pi_0(\Lambda(M)) \cong \mathbb{Z}$ , the homotopy classes of loops being specified precisely by their winding numbers. Since  $T^1$  is a  $K(\pi, 1)$ , Theorem 13 says that each component of relative loop space is also a  $K(\pi, 1)$  with a fundamental group isomorphic to  $\mathbb{Z}$ , and therefore has the homotopy type of a circle.

Now consider  $\Lambda^g(T^1)$  where  $g: T^1 \rightarrow T^1$  is a reflection. Choose one of the two fixed points of the reflection to be the base point  $m$ . We may choose  $\omega$  to be the trivial path from  $m$  to  $g \cdot m$ . Then for each  $\alpha \in \pi_1(T^1, m) \cong \mathbb{Z}$  we have  $\alpha_g = -\alpha$  and so the “ $g$ -twisted action” (46) is the translation

$$\alpha \cdot \beta = \beta - 2\alpha. \quad (47)$$

This has two orbits,  $\overline{\pi_1(T^1)^g} \cong \mathbb{Z}_2$ , and the isotropy subgroups are trivial. It follows from Theorems 12 and 13 that the space of relative loops  $\Lambda^g(T^1)$  has two components, each of which is contractible.

**Numerical Studies** In many interesting problems, typically in celestial mechanics, the action functional is bounded from below and therefore the expected critical points are minimizers. In recent years, in connection with the discovery of the so-called *choreographic periodic orbits*, Simó [52] developed an algorithm to study how to perform a numerical minimization of the action in classes of loops with a definite symmetry type. The idea is based on the description of the orbit in terms of its Fourier coefficients and defining the action  $\mathcal{A}_L[\cdot]$  as a function on the Fourier space. The action  $\mathcal{A}_L[\cdot]$  is then minimized on the space of Fourier coefficients. The interested reader should consult ► *n-Body Problem and Choreographies* in this encyclopedia. It would be very interesting to generalize this method to different actions and to see whether one can impose constraints not only on the symmetry type but also on the topology of the space of loops.

### Hamiltonian Viewpoint

The Hamilton principle gives a variational characterization to the Hamiltonian equation. For Hamiltonian systems in  $\mathbb{R}^{2n}$  the formulation of the principle is very simple. Let  $H: \mathbb{R}^{2n} \rightarrow \mathbb{R}$  be the Hamiltonian function, then the action functional is

**Definition 25 (Action functional in Hamiltonian form)**

$$\mathcal{A}_H[\gamma] \doteq \int_0^T dt \sum_{i=1}^n p_i(t) \dot{q}_i(t) - \int_0^T dt H(q(t), p(t)), \quad (48)$$

where  $\gamma(t) = (q(t), p(t))$ .

It is worth recalling that if the Lagrangian function is hyperregular, then the system can be transformed through the Legendre transform into a Hamiltonian system on the cotangent bundle of the configuration space. In any case, given  $\mathcal{A}_H[\cdot]$  one can show [1,5]

**Proposition 9** *Let  $H$  a smooth Hamiltonian function on  $\mathbb{R}^{2n}$ . Let  $\mathcal{A}_H$  be defined over  $C^2([0, T], \mathbb{R}^{2n})$ , then*

$$D\mathcal{A}_H[q, p](v) = 0 \quad \text{for every } v(\cdot) \in C^2([0, T], \mathbb{R}^{2n})$$

is equivalent to Hamiltonian equations with  $q(0) = q(T)$ ,  $p(0) = p(T)$ .

Recall that in the Hamiltonian context  $p$  and  $q$  are independent variables and therefore to prove the preceding proposition it is necessary to compute the variations with respect to  $q$ 's and  $p$ 's independently. From the analytical point of view the functional  $\mathcal{A}_H[\cdot]$  is a difficult object to study since it is unbounded from below and above, namely, it is *indefinite* [33]. It is not difficult to give an example, consider  $n = 1$  and a Hamiltonian like  $H = p^2/2$ . For indefinite functionals a whole theory has been developed to apply the *mountain pass* theorem [47]. There is also a generalization of the Legendre transform [21]. The new transform [21] can be applied directly to the action functional  $\mathcal{A}_H[\cdot]$  rather than to  $H$ . This allows us to work with a convex functional. We do not enter into the details but we want to recall one result which describes the conditions for the existence of at least  $n$  periodic orbits.

**Theorem 14 ([21])** *Suppose that  $H \in C^1(\mathbb{R}^{2n}, \mathbb{R})$  and for some  $\beta > 0$ , assume that  $C = \{z \in \mathbb{R}^{2n} : H(z) \leq \beta\}$  is strictly convex, with boundary  $\partial C = \{z \in \mathbb{R}^{2n} : H(z) = \beta\}$  satisfying  $\langle z, \nabla H(z) \rangle > 0$  for all  $z \in \partial C$ . Suppose that for  $r, R > 0$  with  $r < R < \sqrt{2}r$  there are two balls  $B_r(0), B_R(0)$  centered at the origin of  $\mathbb{R}^{2n}$  such that  $B_r(0) \subset \text{allowbreak} C \subset B_R(0)$ , then there are at least  $n$*

*distinct periodic solutions on  $\partial C$  of the Hamilton system  $\dot{z} = J \nabla H(z)$ .*

The functional  $\mathcal{A}_H[\cdot]$  has been defined for Hamiltonian systems on  $\mathbb{R}^{2n}$ , but it admits a generalization to symplectic manifolds which are not tangent bundles:

**Definition 26** Let  $(P, \omega)$  be a symplectic manifold and  $H: P \rightarrow \mathbb{R}$  a smooth Hamiltonian function, let  $\gamma: P \rightarrow \mathbb{R}$  be a closed path which is the boundary of a two-dimensional connected region  $\Sigma$ , then

$$\mathcal{A}_H[\gamma] \doteq \int_{\Sigma} \omega - \int_0^T dt H(z(t)). \quad (49)$$

The functional (49) is a very interesting object. It is naturally defined on closed paths which bound two-dimensional regions. The functional can be defined on "paths" but then it would become multivalued. In fact in  $\mathcal{M}$  one has to introduce a 1-form  $\alpha$  such that  $\omega = d\alpha$ . The 1-form  $\alpha$  is not unique and depends on the co-homology of  $\mathcal{M}$ . Although  $\mathcal{A}_H$  is multivalued the differential  $D\mathcal{A}_H$  is single-valued [34]. Let  $\varphi(s)$  be a finite variation with  $\frac{d\varphi(s)}{ds}|_{s=0} = \xi$ , then

$$\begin{aligned} D\mathcal{A}_H[z](\xi) &= \left. \frac{d\mathcal{A}_H[\varphi(s)]}{ds} \right|_{s=0} \\ &= \int_{\Sigma} L_{\xi}(\omega) - \int_0^T dt \langle \nabla H(z), \xi \rangle, \end{aligned}$$

where  $L_{\xi}$  is the Lie derivative with respect to  $\xi$ . Now since  $d\omega = 0$  one can show that

$$D\mathcal{A}_H[z](\xi) = \int_0^T i_{\xi} i_{X_H} \omega - \int_0^T dt \langle \nabla H(z), \xi \rangle,$$

which is single-valued. The theory for such multivalued functionals has been developed by many authors; here we would like to cite [22,43,56]. Note that functionals of the form (49) can result from *symmetry reduction*. In fact as shown in [34] a Lagrangian system with a non-Abelian symmetry  $G$  has a reduced dynamics determined by a variational principle of the form

$$\mathcal{A}_R[\gamma] \doteq \int_0^T dt R(q(t), \dot{q}(t)) - \int_{\Sigma} \beta_{\mu}(q(t), \dot{q}(t)), \quad (50)$$

where  $R$  is the so-called *Routhian* and  $\beta_{\mu}$  is a 2-form dependent on the conserved momentum  $\mu$ . The construction of the Routhian and of the reduced action  $\mathcal{A}_R[\cdot]$  can be found in [34].

### Fixed-Energy Problem, Hill's Region

Let us consider Hamiltonian systems  $(\mathcal{P}, \omega, H)$  where the phase space is a cotangent bundle  $\mathcal{P} = T^*\mathcal{M} \simeq \mathbb{R}^{2n}$ . The symplectic form is then given by  $\omega = d\theta$ , where  $\theta$  is the canonical form, and the Hamiltonian is

$$H(p, q) = \frac{1}{2} \langle p, p \rangle + V(q) \quad (51)$$

$\langle \cdot, \cdot \rangle$  is a Riemannian metric on  $T_q^*\mathcal{M}$ . The Hamiltonian flow preserves the Hamiltonian function; therefore, a natural problem is to restrict the dynamics to the manifold defined by a fixed constant value of the Hamiltonian that physically corresponds to fixing the energy. Letting  $E$  be the energy value, one can define this natural constraint by constructing the following submanifold of the phase space  $\mathcal{P}$ :

$$\Sigma_E = \{(p, q) \in \mathcal{P} : H(p, q) = E\}.$$

From  $\Sigma_E$  it is possible to construct a new manifold that corresponds to the image of the projection of  $\Sigma_E$  onto the configuration space  $\mathcal{M}$ . To obtain such a projection it is sufficient to look at (51) and observe that the norm of  $p$  on  $T_q^*\mathcal{M}$  cannot be negative. From this, the new space, *Hill's region*, turns out to be defined as follows:

#### Definition 27 (Hill's region)

$$P_E \doteq \{q \in \mathcal{M} : E - V(q) \geq 0\}$$

**Remark 12** The manifold  $P_E$  depends on the values of  $E$  and may have boundaries. For instance, consider the Hamiltonian

$$H = \frac{1}{2} \|p\|^2 - \frac{\lambda}{2} \|q\|^2 + \frac{1}{4} \|q\|^4 \text{ with } \lambda > 0,$$

which has a Hill region defined by

$$P_E \doteq \left\{ q \in \mathcal{M} : E + \frac{\lambda}{2} \|q\|^2 - \frac{1}{4} \|q\|^4 \geq 0 \right\}.$$

First note that  $P_E$  is a subset of  $\mathbb{R}^2$  and that there are the following cases:

- $P_E = \emptyset$  for  $E < -\lambda^2/4$ ,
- $P_E$  is a circle for  $E = -\lambda^2/4$ ,
- $P_E$  is an annulus for  $-\lambda^2/4 < E < 0$ ,
- $P_E$  is a disk minus its center for  $E = 0$ ,
- $P_E$  is a disk for  $E > 0$ .

Note that if  $P_E$  is not empty, then the boundary of  $\partial P_E$  is given by

$$\partial P_E \doteq \left\{ q \in \mathcal{M} : E + \frac{\lambda}{2} \|q\|^2 - \frac{1}{4} \|q\|^4 = 0 \right\}$$

and is not empty. The boundary corresponds to the set of points where all momenta  $p$  vanish.

The Hill region has a topology which changes according to the values of the energy; hence, it is a natural problem to search for periodic orbits in it. What are the possible orbits in  $P_E$ ? Assuming there is a generic Hill region with a non-empty boundary, there are two possible types of orbits:

- Orbits joining two points of the boundary, the *brake orbits*,
- Orbits not intersecting the boundary, the *internal periodic orbits*.

In general given a Hamiltonian system with a nonempty Hill region  $P_E$  we define

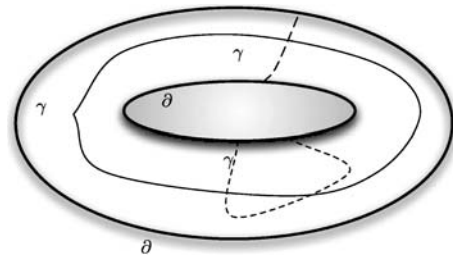
**Definition 28 (Brake orbit)** A solution  $q(\cdot)$  of the Hamilton equation is a *brake orbit* of period  $T$  if  $q(t+T) = q(t)$ ,  $q(t) \in P_E \setminus \partial P_E$  for all  $t \in (0, T/2)$  and  $q(0) \in \partial P_E$ ,  $q(T/2) \in \partial P_E$ .

**Definition 29 (Internal periodic orbit)** A solution  $q(\cdot)$  of the Hamilton equation is an *internal periodic orbit* of period  $T$  if  $q(t+T) = q(t)$  and  $q(t) \in P_E \setminus \partial P_E$  for all  $t$ .

**Remark 13** In principle one could think that brake orbits could have more than two intersections with the boundary  $\partial P_E$ . This is not possible in systems with Hamiltonian (51) because the velocity on  $\partial P_E$  is zero and the Hamiltonian equations are “reversible”, that is, if  $q(t)$  is a solution then  $q(-t)$  is a solution. These two properties imply that any solution  $q(t)$  such that  $q(t_i) \in \partial P_E$  follows the trajectory  $q(-t)$  for  $t > t_i$  with reversed velocity [24,30].

An example of Hill's region with brake orbits and internal periodic orbit is given in Fig. 4.

**Jacobi Metric and Tonelli Functional** The general approach to study the periodic orbits of type A or B is to use



Periodic Orbits of Hamiltonian Systems, Figure 4

Example of a compact nonsimply connected Hill region with two brake orbits  $\gamma_{\text{brake}}$  and one internal orbit  $\gamma_{\text{internal}}$  that cannot be deformed into a point

a variational principle which is naturally defined on paths in  $P_E$ . Let us consider

**Definition 30 (Jacobi metric)**

$$\mathcal{J}_E[q] = \frac{1}{2} \int_0^1 ds (E - V(q(s))) \left\| \frac{dq(s)}{ds} \right\|^2. \quad (52)$$

The domain of  $\mathcal{J}_E[\cdot]$  is  $\Lambda(P_E) \doteq \{q(\cdot) \in H^1([0, 1], \mathbb{R}^n) : q(s+1) = q(s) \text{ } q(s) \in P_E\}$ , and the following result holds:

**Proposition 10** *Let  $q(\cdot)$  be a path in  $\Lambda(P_E)$  where  $\mathcal{J}_E[\cdot]$  is differentiable and let  $q(\cdot) + \epsilon v(\cdot) \in \Lambda(P_E \setminus \partial P_E)$  for all  $\epsilon$  sufficiently small, then*

$$D\mathcal{J}_E[q](v) = 0 \quad \text{for all } v(\cdot) \in T\Lambda(P_E)$$

is equivalent to

$$\frac{dq(t)}{dt} = -\nabla V(q(t))$$

after a time rescaling defined by

$$\frac{dt}{ds} = \left\| \frac{dq(s)}{ds} \right\| \frac{1}{\sqrt{E - V(q(s))}}.$$

The proof of this proposition is quite standard and can be found in [1,8,11,30,51]. Let us observe that in the regions where  $E - V(q) > 0$  the functional  $\mathcal{J}_E[\cdot]$  can be derived from the Riemannian metric

$$d^2\ell = (E - V(q)) \sum_{i,j} \delta_{ij} dq_i \otimes dq_j \quad (53)$$

and indeed from this its name originated. From the Jacobi metric a length can be defined

$$\ell[q] \doteq \int_0^1 ds \sqrt{(E - V(q(s))) \|\dot{q}(s)\|^2}.$$

The Jacobi metric (53) transformed the study of the classical Newton equations into the study of the geodesics on  $P_E$ . It is crucial to note that the Jacobi metric is vanishing on  $\partial P_E$  and therefore cannot be complete in  $P_E$  whenever the boundary is not empty. There is another functional that can be used to study periodic orbits in  $P_E$ . This is defined as follows:

**Definition 31 (Tonelli functional)**

$$\mathcal{T}_E[q] = \frac{1}{2} \int_0^1 ds (E - V(q(s))) \int_0^1 ds \left\| \frac{dq(s)}{ds} \right\|^2.$$

Tonelli's functional is a product of two integrals and therefore does not have a natural geometrical interpretation like

(53). One can observe that (by Schwartz inequality) the Tonelli functional is bounded from below by Jacobi length:  $(\ell[q])^2 \leq 2\mathcal{T}_E[q]$ . Note that also  $\mathcal{T}_E[\cdot]$  vanishes on  $\partial P_E$ . The functional  $\mathcal{T}_E[\cdot]$  is defined on  $\Lambda(P_E)$  and provides another possible variational description of Newton's equations. Indeed one can easily show [2].

**Proposition 11** *Let  $q(\cdot)$  be a path in  $\Lambda(P_E)$  where  $\mathcal{T}_E[\cdot]$  is differentiable and let  $q(\cdot) + \epsilon v(\cdot) \in \Lambda(P_E \setminus \partial P_E)$  for all  $\epsilon$  sufficiently small, then*

$$D\mathcal{T}_E[q](v) = 0 \quad \text{for all } v(\cdot) \in T\Lambda(P_E)$$

is equivalent to

$$\frac{dq(t)}{dt} = -\nabla V(q(t))$$

after time rescaling defined by

$$\left( \frac{dt}{ds} \right)^2 = \frac{\int_0^1 ds \left\| \frac{dq(s)}{ds} \right\|^2}{\int_0^1 ds (E - V(q(s)))}.$$

**Remark 14** Observe that in both the Jacobi and the Tonelli formulation there is a reparameterization of the time and therefore one can always restrict the consideration to trajectories with unit period.

In variational methods we aim to show that the critical set is not empty and, if possible, to estimate its cardinality. This is certainly affected by the topology of the space, where the paths are defined. Let us now consider the Jacobi metric  $\mathcal{J}_E[\cdot]$ . This functional depends on the energy  $E$ , which affects the properties of  $P_E$  and  $\Lambda(P_E)$ . There are four possible cases:

- $P_E$  is compact and  $\partial P_E = \emptyset$ ,
- $P_E$  is compact and  $\partial P_E \neq \emptyset$ ,
- $P_E$  is not compact and  $\partial P_E = \emptyset$ ,
- $P_E$  is not compact and  $\partial P_E \neq \emptyset$ .

In the next section we shall illustrate some results regarding the preceding four cases. We shall see that the main approaches have much in common with Riemannian geometry. For more details on variational methods the reader is encouraged to consult [2,15,28,33,53].

**Orbits in Compact Hill's Regions Without a Boundary**

Let us consider a region  $P_E$  without a boundary for  $E > \sup_{q \in \mathcal{M}} V(q)$ . If  $\mathcal{M}$  is compact, then the previous condition can be satisfied for some finite  $E$ . In this case

$P_E \simeq \mathcal{M}$  and the Jacobi metric  $J_E[\cdot]$  becomes a Riemannian metric on  $\mathcal{M}$ ; therefore, the problem of periodic orbits in  $\mathcal{M}$  translates into the problem of closed geodesics in the Riemannian manifold  $\mathcal{M}$ . This is solved by

**Theorem 15 (Lusternik and Fet)** *Each compact Riemannian manifold contains a closed geodesic.*

For a proof see [28,29]. The main tool for proving the theorem is the so-called *curve shortening*. The manifold  $P_E$  is Riemannian with metric (53). A base for the topology is given by  $B_r(q) = \{q' \in \mathcal{M} : d(q, q') < r\}$ , where

$$d(q, q') = \inf \{ \ell[\gamma], \gamma(\cdot) \text{ is a piecewise smooth curve such that } \gamma(0) = q \text{ and } \gamma(1) = q' \}.$$

**Remark 15** The reader may observe that if  $P_E$  has a non-empty boundary then  $d(\cdot, \cdot)$  turns into a pseudometric because  $d(q, q') = 0$  for all  $q, q' \in \partial P_E$ . Certainly  $d(\cdot, \cdot)$  is still a metric restricted to the open set  $P_E \setminus \partial P_E$ .

Let us now give a sketch of the curve-shortening method. A standard result in Riemannian geometry guarantees that given a point  $q_0$  and a neighborhood  $B_\delta(q_0)$ , there exists  $\delta$  such that any point in  $\partial B_\delta(q_0)$  can be joined to  $q_0$  by a unique geodesic [28]. Since  $\mathcal{M}$  is compact one can use a finite family of such neighborhoods to join any two points  $q_0, q_1 \in \mathcal{M}$  with a piecewise geodesic  $\gamma(\cdot)$ . Now consider on  $\mathcal{M}$  the space of closed paths of class  $H^1([0, 1], \mathcal{M})$  and consider a sequence  $0 \leq t_0^{(k)} < t_1^{(k)} < \dots < t_{n-1}^{(k)} < t_n^{(k)} \leq 1$  such that  $t_{i+1} - t_i < \delta^2/(2c)$ . Take a curve  $\gamma(\cdot) \in H^1([0, 1], \mathcal{M})$  with  $J_E[\gamma] \leq c$  then define

$$S^{(k)}(\gamma) = \begin{cases} \text{is a piecewise geodesic curve} \\ \text{where } S^{(k)}(\gamma)(t) \text{ is restricted to } t_i^{(k)}, t_{i+1}^{(k)}, \\ \text{is a geodesic joining } \gamma t_i^{(k)} \text{ to } \gamma t_{i+1}^{(k)} \\ \text{for } i = 1 \dots n. \end{cases}$$

Now clearly  $J_E[S^{(k)}(\gamma)] \leq J_E[\gamma]$ ,  $\ell[S^{(k)}(\gamma)] \leq \ell[\gamma]$ . The map  $S^{(k+1)}(\cdot)$  is constructed by taking a new partition of  $[0, 1]$  such that  $t_i^{(k)} < t_i^{(k+1)} < t_{i+1}^{(k)}$ . Since one can easily verify

$$\ell[S^{(k+1)}(\gamma)] \leq \ell[S^{(k)}(\gamma)] \quad (54)$$

the iteration of  $S^{(k)}(\cdot)$  is called *curve shortening*. In [24,28], it is shown that  $S^{(k)}(\gamma)$  converges uniformly to a geodesic. The limit could be just a point curve. To show that this cannot be the case one uses the fact that on a compact manifold of dimension  $n$  there is always a map  $h: S^d \mapsto \mathcal{M}$  (with  $1 \leq d \leq n$ ) which is not homotopic to a constant map.

There have been many generalizations of Lusternik-Fet theorem and the reader is invited to refer to [29].

### Brake Orbits in Hill's Regions with a Boundary

Let  $P_E$  be a region with boundary for  $\inf_{q \in \mathcal{M}} V(q) < E < \inf_{q \in \mathcal{M}} V(q)$ . In this case there is the following result:

**Theorem 16 ([11])** *Suppose that  $P_E$  is compact and there are no equilibrium positions in  $\partial P_E$ . Then the number of brake orbits in  $P_E$  is at least equal to the number of generators of  $\pi_1(P_E \setminus \partial P_E)$ .*

Since the Jacobi metric is singular on  $\partial P_E$  it is necessary to analyze the geodesic motion near the boundary. This type of analysis goes back to earlier works [9,10,51]. The main point is to construct a new region  $P_{E-\epsilon} \subset P_E$  on which  $J_E[\cdot]$  is positive-definite. The next step is to construct a geodesic joining two points on the new boundary  $\partial P_{E-\epsilon}$ . Finally the construction on  $P_{E-\epsilon}$  is used to show that the geodesic  $q^\epsilon(\cdot)$  becomes a brake orbit in the limit  $\epsilon \rightarrow 0$ . In the case of noncompact  $P_E$ , the existence of a brake orbit was proved in [44]. This result is based on two assumptions:

- (i)  $\partial P_E$  is not empty and it is formed by two connected components  $A_1$  and  $A_2$  such that if  $x_n \in A_1$  and  $y_n \in A_2$  with  $\|x_n\| \rightarrow \infty$ ,  $\|y_n\| \rightarrow \infty$  then  $\|x_n - y_n\| \rightarrow \infty$ .
- (ii) Let  $R_\delta = \{q \in \mathcal{M} : E - \delta < V(q) < E\}$  for  $\delta > 0$ . If either  $A_1$  or  $A_2$  is not compact, there is a number  $r > 0$  with the following property: if  $r_0 > r$  then there exist  $r_1 > r_0$  and  $\delta > \delta^* > 0$  such that for every  $q \in R_\delta \cap \{q : \|q\| > r_1\}$  and  $(q, p) \in H^{-1}(E)$  implies that the Hamiltonian flow  $\phi(t, q, p)$  stays in  $\{q : \|q\| > r_1\}$  for all  $t \geq 0$  where defined.

**Theorem 17 ([44])** *Suppose that  $P_E$  is connected,  $\inf_{q \in \partial P_E} \|\nabla V(q)\| > 0$  and conditions i and ii hold true. Then there exists a periodic solution which is a brake orbit.*

In this result the lack of compactness of  $P_E$  does not allow us to use the curve shortening. For this reason in [44] the direct minimization was employed. First a new region  $P_{E-\delta} \subset P_E$  is defined for  $\delta > 0$ . Such region now has a boundary with two different connected components,  $A_1^\delta$  and  $A_2^\delta$ , respectively. In [44] the following minimization problem was studied

$$\min J_E[q] \text{ where } q(\cdot) \in H^1$$

$$\text{and boundary conditions } q(0) \in A_1^\delta, q(1) \in A_1^\delta.$$



Condition i is able to control the lack of compactness of  $P_E$  and to show that the minimization problem admits a solution  $q^\delta(\cdot)$ . Condition ii allows us to take the limit  $\delta \rightarrow 0$  to obtain a brake orbit in  $P_E$ .

### Orbits in Closed, Nonsimply Connected Hill's Regions with Boundaries

Let us now consider a general Hill region which is not necessarily compact, with a boundary and with nontrivial homotopy group  $\pi_1(P_E)$ . The natural problem is to determine under which condition it is possible to prove the existence of an internal periodic orbit within a specific homotopy class of  $P_E$ . A partial answer is given by

**Theorem 18 ([6,7])** *Suppose that  $P_E$  is closed and bounded and that  $\nabla V(q) \neq 0$  for all  $q \in \partial P_E$ . Then there exists a periodic orbit  $q(s) \in P_E$  for all  $s$ . The orbit  $q(\cdot)$  can be either a brake orbit or an internal periodic orbit.*

As we have already seen, the Jacobi metric (but also the Tonelli functional) is degenerate on the boundary  $\partial P_E$ , i.e.,  $J_E[\gamma]$  is identically zero on every closed path  $\gamma(\cdot)$  such that  $\gamma(s) \in \partial P_E$  for every  $s \in [0, 1]$ . In order to avoid this problem, in [6,7], a modified functional was introduced:

$$\mathcal{J}^\epsilon[q] \doteq J_E[q] + \int_0^1 ds U^\epsilon[q(s)].$$

The functional  $\mathcal{J}^\epsilon[\cdot]$  is called *penalized*. If  $q_n(\cdot)$  is a sequence of curves in  $\overline{\Lambda(P_E)}$  (the closure of  $\Lambda(P_E)$ ) weakly converging to a curve  $q^*(\cdot)$  intersecting the boundary, then the form of  $U^\epsilon(\cdot)$  implies that  $\mathcal{J}^\epsilon[q_n] \rightarrow +\infty$  whenever  $\epsilon > 0$ . This type of penalization is similar to the so-called *strong force* potential used in the  $N$ -body problem (see [2,26] and also ► [n-Body Problem and Choreographies](#)). In this case there are two kinds of difficulties:  $P_E$  is no longer compact and the boundary  $\partial P_E$  is not empty. The strategy is as follows: prove that the critical level sets of  $\mathcal{J}^\epsilon[\cdot]$  are precompact. This is obtained by a generalization of the Palais–Smale condition. This prevents the sequences from converging on the boundary, and it is realized by taking a functional  $\rho: C^1(\Lambda(P_E)) \rightarrow \mathbb{R}_+$  such that if  $q_n(\cdot)$  converges to a path intersecting the boundary, then  $\rho(q_n) \rightarrow +\infty$ . This allows us to introduce:

### Definition 32 (Weighted Palais–Smale condition [6])

The action  $\mathcal{J}^\epsilon[\cdot]$  satisfies the weighted Palais–Smale condition if any sequence  $q_n(\cdot)$  fulfills one of the following alternatives:

1.  $\rho(q_n)$  and  $\mathcal{J}^\epsilon[q_n]$  are bounded and  $D\mathcal{J}^\epsilon[q_n] \rightarrow 0$  and  $q_n$  has a convergent subsequence,
2.  $\mathcal{J}^\epsilon[q_n]$  is convergent and  $\rho(q_n) \rightarrow +\infty$  and there exists  $\nu > 0$  such that  $\|D\mathcal{J}^\epsilon[q_n]\| \geq \nu \|D\rho(q_n)\|$  for  $n$  sufficiently large.

The verification of the weighted Palais–Smale condition guarantees that the set  $K_c \cap \{q: \rho(q) \leq M\}$  is compact for every  $M > 0$ . The next step in the proof is to construct a mini-max structure [47] that allows us to identify the critical values. This is achieved by constructing two manifolds  $Q$  and  $S$  such that:

- (i)  $S \cap \partial Q = \emptyset$ ,
- (ii) If  $u \in C^0(\overline{Q}, \Lambda(P_E))$  such that  $u(q(\cdot)) = q(\cdot)$  for every  $q(\cdot) \in \partial Q$  then  $h(\overline{Q}) \cap S \neq \emptyset$ .

The manifolds  $Q$  and  $S$  are defined such that the critical level

$$c = \inf_{u \in U} \sup_{q \in Q} \mathcal{J}^\epsilon(h(q))$$

is finite. Here  $U = \{u \in C^0(\overline{Q}, \overline{\Lambda(P_E)}): u(q(\cdot)) = q(\cdot) \text{ if } \mathcal{J}^\epsilon[q] \leq 0\}$ . Then the weighted Palais–Smale condition allows us to construct a gradient flow and to prove the existence of a critical point  $q^\epsilon(\cdot)$ . In [7] it is shown also that there exists  $\alpha \leq \beta$  independent of  $\epsilon$  such that

$$\alpha \leq \mathcal{J}^\epsilon[q^\epsilon] \leq \beta.$$

This uniform estimates make it possible to prove that for  $\epsilon \rightarrow 0$  the path  $q^\epsilon(\cdot)$  converges uniformly to a closed path  $q^0(\cdot)$  such that the set  $I \doteq \{t \in [0, 1]: q^0(t) \in P_E \setminus \partial P_E\}$  is not empty. Since  $q^0(\cdot)$  is continuous then the interval  $I$  has connected components. Note that  $q^0(\cdot)$  has the same homotopy type as  $q^\epsilon(\cdot)$ , but in general this is no longer true for the solution of the equations of motion. In fact  $q^0(\cdot)$  is a solution only on the connected component of  $I$ . Let  $\Delta$  be one connected component of  $I$ . On  $\Delta$  the path  $q^0(\cdot)$  solves the equations of motion. Therefore, the topological characterization of the periodic orbit depends on  $\Delta$ . If  $\Delta \subset [0, 1]$  then  $q^0(\cdot)$  is a *brake orbit*, if  $\Delta = [0, 1]$  then  $q^0(\cdot)$  is an *internal periodic orbit*. The problem of finding internal periodic orbits in any prescribed homotopy class is still open.

### Continuation of Periodic Orbits as Critical Points

Variational methods can also be very useful to study the problem of continuation of periodic orbits. Here we restrict ourselves to a few examples.

**Lagrangian Variational Principle** Let us consider a dynamical system in  $\mathbb{R}^n$  that can be written in the Lagrangian form:

$$L(v_q, q) = \frac{1}{2} \|v_q\|^2 - V(q).$$

Let us assume

- (i) There exists  $a > 0$  such that  $V(\lambda q) = \lambda^{-a} V(q)$  for any  $\lambda > 0$ ,
- (ii) There exists a 1-periodic orbit  $q_1(\cdot)$ .

The orbit  $q_1(\cdot)$  is a critical point of  $\mathcal{A}_L[q] = \int_0^1 ds L(\dot{q}(s), q(s))$ . Now let us consider another potential term  $W(q)$  such that

$$W(\lambda q) = \lambda^{-b} W(q) \quad \text{with } b > a.$$

Let us look for periodic orbits of period  $T$  for the system whose Lagrangian is

$$L(v_q, q) = \frac{1}{2} \|v_q\|^2 - V(q) - W(q).$$

This suggests searching for critical points  $x(\cdot)$  of the functional

$$\mathcal{A}_L^T[x] = \int_0^T dt \left[ \frac{\|\dot{x}(t)\|^2}{2} - V(x(t)) - W(x(t)) \right]. \quad (55)$$

The scaling properties of  $V$  and  $W$  can be used to construct a perturbation argument. Define

$$x(t) = T^{-p} q(t/T), \quad \text{where } q(s) \text{ is defined for } s \in [0, 1], \text{ and } p = 2/(a+2). \quad (56)$$

The dynamical equations for  $x(\cdot)$  are equivalent to the Lagrangian equation for

$$\mathcal{A}_L[q; \epsilon] = \int_0^1 ds \left[ \frac{\|\dot{q}(s)\|^2}{2} - V(q(s)) - \epsilon W(q(s)) \right], \quad (57)$$

where

$$\epsilon \doteq T^{-\frac{2(b-a)}{a+2}}.$$

The scaling properties of the potential allow us to transform the given problem into a perturbation problem. Note that there is the following correspondence: a small

perturbation for the scaled action (57) corresponds to large periods for (55). Now  $\mathcal{A}_L[\cdot, 0] = \mathcal{A}_L^1[\cdot]$  and in general the critical point  $q_1(\cdot)$  is not isolated and therefore  $D^2 \mathcal{A}_L[q_1, 0]$  has some degeneracy. If the manifold of critical points is degenerate along normal directions (*normal degeneracy*), then it is possible to continue the critical point  $q_1(\cdot)$  into a  $T$ -periodic orbit by decomposing the continuation procedure. This approach is the so-called *Liapunov-Schmidt reduction*; it is used to generalize the IFT to situations where it cannot be applied directly. An example will be illustrated in the last part of this section in the study of Hamiltonian systems. For details about normal degeneracy the reader could consult [15].

**Hamiltonian Variational Principle** Hamiltonian equations can be thought of as a vector field on a space of loops. The geometrical construction was described in [56]. Let us consider the space  $C^1([0, 1], \mathbb{R}^{2n})$  of differentiable loops. We can define

$$\mathcal{X}(z) \doteq \nu \frac{dz(s)}{ds} - J \nabla H(z(s)), \quad z(\cdot) \in C^1([0, 1], \mathbb{R}^{2n}). \quad (58)$$

The zero set of  $\mathcal{X}(z)$  is formed by loops  $z(\cdot)$  such that

$$\nu \frac{dz(s)}{ds} = J \nabla H(z(s)).$$

This corresponds to a periodic orbit  $z(\cdot)$  with period  $2\pi/\nu$ . It turns out that  $\mathcal{X}(\cdot)$  is a Hamiltonian vector field whose Hamiltonian on  $C^1([0, 1], \mathbb{R}^{2n})$  [25,56] is given by

$$\mathcal{H}[z] = \frac{1}{2\pi} \int_0^{2\pi} ds \{ \langle \nu J \dot{z}(s), z(s) \rangle - H(z(s)) \} \quad (59)$$

and with symplectic form

$$\Omega[z, w] = \frac{1}{2\pi} \int_0^{2\pi} ds \langle Jz(s), w(s) \rangle. \quad (60)$$

Indeed a simple calculation shows that

$$\Omega(\mathcal{X}(z), w) = d\mathcal{H}[z](w).$$

Let us now suppose there is a periodic orbit  $z_0(\cdot)$ . We want to illustrate how to study the continuation and bifurcation when the Hamiltonian is perturbed. In order to use the formulation in the loop space we can introduce the Liapunov-Schmidt reduction. In fact, in general, Hamiltonian vector fields at a periodic orbit have a linearization

with a nonempty kernel. The construction we now present is a very well known approach in the infinite-dimensional setting.

**Liapunov–Schmidt Reduction for Hamiltonian Systems** [3,25] Let  $(\mathcal{H}, \mathcal{P}, \Omega)$  be a Hamiltonian system with  $\mathcal{P} = C^1([0, 1], \mathbb{R}^{2n})$ . Let  $X: \mathcal{P} \rightarrow \mathcal{Y}$  be the Hamiltonian vector field such that  $z_0(\cdot)$  is a zero of  $X(\cdot)$  and

- (i)  $L = DX(z_0)$  is Fredholm,
- (ii)  $J(\ker(L)) = \ker(L)$ .

Condition i implies that it is possible to make the following decomposition

1.  $\mathcal{P} = \ker(L) \oplus W$ ,  $W$  closed subspace,
2.  $\mathcal{Y} = \text{rank}(L) \oplus N$ ,  $N$  closed subspace.

In fact any  $z$  can be written as  $z = k + w$  with  $k \in \ker(L)$  and  $w \in W$ . This allows us to reduce  $X(z) = 0$  to solving the following system of equations:

$$\begin{cases} \Pi X(k + w) = 0 \in \text{rank}(L) \\ (\mathbf{I} - \Pi) X(k + w) = 0 \in N \end{cases} \quad (61)$$

where  $\Pi: \mathcal{Y} \mapsto \text{rank}(L)$ . The first equation in (61) can be solved with respect to  $w$  by the IFT. The second equation becomes the so-called *bifurcation* equation:

$$X_g(k) = (\mathbf{I} - \Pi)X(k + w(k)) = 0. \quad (62)$$

Note that since  $L$  is Fredholm  $\dim \ker(L) < \infty$ , (62) is a finite-dimensional problem. Now condition ii implies that  $X_g$  can be thought of as a map from  $\ker(L)$  into itself. Furthermore this allows us to show that  $X_g$  is a Hamiltonian vector field with Hamiltonian

$$g(k) = \mathcal{H}(k + w(k))$$

and symplectic form  $\Omega(\cdot, \cdot)$ . If condition ii does not hold a further condition has to be included in order to guarantee that  $X_g$  is a Hamiltonian vector field. The bifurcation equation becomes therefore

$$\Omega(X_g(k), u) - d\mathcal{H}(k + w(k))(u) = 0 \quad \forall u. \quad (63)$$

In [56] the loop space approach and the reduction were used to show that if a Hamiltonian system has a manifold  $\Sigma$  of periodic orbits whose tangent space coincides with the kernel of  $D^2\mathcal{A}_H$  (*nondegeneracy condition*) then small perturbations of the Hamiltonian cannot destroy all the periodic orbits:

**Theorem 19** ([56]) *Let  $\Sigma$  be a compact, nondegenerate manifold of periodic orbits for a Hamiltonian system  $(\mathcal{P}, \omega, H)$ . Given any neighborhood  $U$  of  $\Sigma$  there exists  $\epsilon_0 > 0$  such that for  $|\epsilon| < \epsilon_0$ , the number of periodic orbits in  $U$  for  $(\mathcal{P}, \omega, H + \epsilon H_1)$  is no less than the Lusternik–Schnirelman category  $\text{Cat}(\Sigma/S^1)$ . If  $\Sigma$  satisfies the additional condition that the algebraic multiplicity of 1 as an eigenvalue of the Poincaré map is uniformly equal to the  $\dim \Sigma$ , then  $\text{Cat}(\Sigma/S^1) \geq (1 + \dim \Sigma)/2$ .*

In [25] Liapunov–Schmidt reduction was used to prove the Liapunov center theorem and also the Hamiltonian Hopf bifurcation.

### Further Directions

This article has focused on periodic orbits in Hamiltonian systems. This problem can be seen as an organizing center in the history of the development of modern mathematics. In fact it is related to many theoretical aspects: bifurcation theory, symmetry reduction, variational methods and topology of closed curves on manifolds. Still, open problems remain, for example, one would like to prove the existence of internal periodic orbits in every homotopy class of a Hill region. Another interesting direction is the analytical study of the action functional that results from symmetry reduction. A further generalization of the Hamiltonian formalism is the study of the so-called multiperiodic patterns. Here is an example. Let  $\phi, p$  be two functions defined on the two-dimensional torus  $\mathbb{T}^2$  and valued in  $\mathbb{R}$ . The functions  $\phi, p$  are 2-periodic because they satisfy  $p(x + \tau_1, y + \tau_2) = p(x, y)$ ,  $\phi(x + \tau_1, y + \tau_2) = \phi(x)$ , where  $(x, y) + (\tau_1, \tau_2) \simeq (x, y)$  in  $\mathbb{T}^2$ . One can pose the problem to solve a couple of partial differential equations of the form

$$\begin{aligned} \nabla^2 \phi + \frac{\partial}{\partial \phi} V(\phi, p, x, y) &= 0 \quad \text{and} \\ \nabla^2 p + \frac{\partial}{\partial p} V(\phi, p, x, y) &= 0. \end{aligned} \quad (64)$$

Here  $V: \mathbb{R}^2 \mapsto \mathbb{R}$  is a smooth potential function. Note that the periodicity of  $\phi$  and  $p$  is partial, that is why the term “multiperiodic patterns” is used. It has been shown that Eqs. (64) admits a description in terms of finite-dimensional multi-symplectic structure and the equations of motion can be cast into a general Hamiltonian variational principle. In fact one can define  $Z = (\phi, u_1, u_2, p)$  defined as functions on  $\mathbb{T}^2$  and the Hamiltonian

$$S(Z, x, y) = \frac{1}{2}(u_1^2 + u_2^2) + V(\phi, p, x, y).$$

Then Eqs. (64) turn out to be equivalent to the variational equation

$$D\mathcal{L}(Z) = 0, \quad (65)$$

where  $\mathcal{L}$  is the action functional

$$\mathcal{L}(Z) = \int_0^{\tau_1} dx \int_0^{\tau_2} dy \left( \frac{1}{2} \left\langle J_1 \frac{\partial Z}{\partial x} + J_2 \frac{\partial Z}{\partial y}, Z \right\rangle - S(Z, x, y) \right) \quad (66)$$

defined on 2-periodic maps  $Z: \mathbb{T}^2 \mapsto \mathbb{R}^4$ . The matrices  $J_1$  and  $J_2$  are two symplectic structures and form a multisymplectic structure. In this example  $J_1$  and  $J_2$  read

$$J_1 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad (67)$$

$$J_2 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

For Eqs. (64) one can pose the problem of finding solutions as “multi-periodic” critical points of the action  $\mathcal{L}$ , which is a generalization of the Hamiltonian action functional (48). An interesting reference for this problem is [13].

## Acknowledgments

The author would like to thank Heinz Hanßmann for his critical and thorough reading of the manuscript and Ferdinand Verhulst for his useful suggestions.

## Bibliography

- Abraham R, Marsden J (1978) Foundations of Mechanics. Benjamin-Cummings, Reading
- Ambrosetti A, Coti V (1994) Zelati Periodic Solutions of singular Lagrangian Systems. Birkhäuser, Boston
- Ambrosetti A, Prodi G (1993) A primer on Nonlinear Analysis. CUP, Cambridge UK
- Arms JM, Cushman R, Gotay MJ (1991) A universal reduction procedure for Hamiltonian group actions. In: Ratiu TS (ed) The Geometry of Hamiltonian Systems. Springer, New York, pp 33–51
- Arnold VI (1989) Mathematical Methods of Classical Mechanics. Springer, New York
- Benci V (1984) Normal modes of a Lagrangian system constrained in a potential well. Annales de l'IHP section C tome 5(1):379–400
- Benci V (1984) Closed geodesics for the Jacobi metric and periodic solutions of prescribed energy of natural Hamiltonian systems. Annales de l'IHP section C tome 5(1):401–412
- Benci V (1986) Periodic solutions for Lagrangian systems on a compact manifold. J Diff Eq 63:135–161
- Birkhoff GD (1927) Dynamical Systems with two degrees of freedom. Trans Am Math Soc 18:199–300
- Birkhoff GD (1927) Dynamical Systems. Colloq. Publ. 9. Amer. Math. Soc., Providence RI
- Bolotin SV, Kozlov VV (1978) Libration in systems with many degrees of freedom. J Appl Math Mech 42:256–261
- Bott R, Tu W (1995) Differential forms in algebraic topology. Springer, New York
- Bridges T (2006) Canonical multi-symplectic structure on the total exterior algebra bundle. Proc R Soc A 462:1531–1551
- Brown KS (1982) Cohomology of groups. Springer, New York
- Chang K (1991) Infinite dimensional Morse theory and multiple solution problems. Birkhäuser, Boston
- Cushman R, Bates L (1997) Global aspects of classical integrable systems. Birkhäuser, Boston
- Cushman R, Sadowski D, Efstathiou K (2005) No polar coordinates. In: Montaldi J, Ratiu T (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge University Press, Cambridge UK
- Dell Antonio G (1995) Agomenti scelti di meccanica SISSA/ISAS Trieste. ref. ILAS/FM-19/1995
- Doedel EJ, Keller HB, Kernvez JP (1991) Numerical analysis and control of bifurcation problems: (I) Bifurcation in finite dimensions. Int J Bifurcat Chaos 3(1):493–520
- Doedel EJ, Keller HB, Kernvez JP (1991) Numerical analysis and control of bifurcation problems: (II) Bifurcation in infinite dimensions. Int J Bifurcat Chaos 4(1):745–772
- Ekeland I (1990) Convexity methods in Hamiltonian mechanics. Springer, Berlin
- Farber M (2004) Topology of Closed One-Forms. AMS, Providence
- Gallavotti G (2001) Quasi periodic motions from Hipparchus to Kolmogorov. Rend Mat Acc Lincei s. 9, 12:125–152
- Gluck H, Ziller W (1983) Existence of periodic motions of conservative systems. In: Bombieri E (ed) Seminar on Minimal Submanifolds. Princeton University Press, Princeton NJ
- Golubitsky M, Marsden JE, Stewart I, Dellnitz M (1995) The constrained Liapunov–Schmidt procedure and periodic orbits. Fields Institute Communications 4. Fields Institute, Toronto, pp 81–127
- Gordon W (1975) Conservative dynamical systems involving strong force. Trans Am Math Soc 204:113–135
- Guillemin V, Sternberg S (1984) Symplectic techniques in Physics. Cambridge University Press, Cambridge UK
- Jost J (1995) Riemannian geometry and geometric analysis. Springer, Berlin
- Klingenberg W (1978) Lectures on closed geodesics. Springer, Berlin
- Kozlov VV (1985) Calculus of variations in the large and classical mechanics. Russ Math Surv 40:37–71
- Kuznetsov YA (2004) Elements of Applied Bifurcation Theory. Springer, Berlin

32. Leimkuhler B, Reich S (2005) *Simulating Hamiltonian Dynamics*. CUP, Cambridge UK
33. Mawhin J, Willhelm M (1990) *Critical point theory and Hamiltonian systems*. Springer, Berlin
34. Marsden J, Ratiu TS, Scheurle J (2000) Reduction theory and the Lagrange–Routh equations. *J Math Phys* 41:3379–3429
35. McCord C, Montaldi J, Roberts M, Sbano L (2003) *Relative Periodic Orbits of Symmetric Lagrangian Systems*. Proc. Equadiff. World Scientific, Singapore
36. Meyer K, Hall GR (1991) *Introduction to Hamiltonian Systems and the N-Body Problem*. Springer, Berlin
37. Meyer K (1999) Periodic solutions of the  $N$ -body problem. *LNM*, vol 1719. Springer, Berlin
38. Montaldi J, Buono PL, Laurent F (2005) Poltz Symmetric Hamiltonian Bifurcations. In: Montaldi J, Ratiu T (eds) *Geometric Mechanics and Symmetry: the Peyresq Lectures*. Cambridge University Press, Cambridge UK
39. Montaldi J, Roberts M, Stewart I (1988) Periodic solutions near equilibria of symmetric Hamiltonian systems. *Phil Trans R Soc Lon A* 325:237–293
40. Montaldi J, Roberts M, Stewart I (1990) Existence of nonlinear normal modes of symmetric Hamiltonian systems. *Nonlinearity* 3:695–730
41. Montaldi J, Roberts M, Stewart I (1990) Stability of nonlinear normal modes of symmetric Hamiltonian systems. *Nonlinearity* 3:731–772
42. Munoz–Almaraz FJ, Freire E, Galan J, Doedel E, Vanderbauwhede (2003) A Continuation of periodic orbits in conservative and Hamiltonian systems. *Physica D* 181:1–38
43. Novikov SP (1982) The Hamiltonian formalism and a multivalued analogue of Morse theory. *Russ Math Surv* 37:1–56
44. Offin D (1987) A Class of Periodic Orbits in Classical Mechanics. *J Diff Equ* 66:9–117
45. Ortega J, Ratiu T (1998) Singular Reduction of Poisson Manifolds. *Lett Math Phys* 46:359–372
46. Poincaré H (1956) *Le Méthodes Nouvelles de la Mécanique Céleste*. Gauthiers-Villars, Paris
47. Rabinowitz PH (1986) *Minimax Methods in Critical Point Theory with Applications to Differential Equations*. In: CBMS Reg. Conf. Ser. No. 56, Amer. Math. Soc., Providence, RI
48. Ratiu T, Sousa Dias E, Sbano L, Terra G, Tudora R (2005) A crash course. In: Montaldi J, Ratiu T (eds) *geometric mechanics in Geometric Mechanics and Symmetry: the Peyresq Lectures*. Cambridge University Press, Cambridge UK
49. Sanders J, Verhulst F, Murdock J (2007) *Averaging Methods in Nonlinear Dynamical Systems*. Applied Mathematical Sciences, vol 59. Springer, Berlin
50. Schmäh T (2007) A cotangent bundle slice theorem. *Diff Geom Appl* 25:101–124
51. Seifert H (1948) Periodische Bewegungen mechanischer Systeme. *Math Z* 51:197–216
52. Simó C (2001) New families of solutions in  $N$ -body problems. In: *European Congress of Mathematics*, vol I (Barcelona, 2000), pp 101–115, Progr. Math., 201. Birkhäuser, Basel
53. Struwe M (1990) *Variational methods*. Springer, Berlin
54. Verhulst F (1990) *Nonlinear Dynamical Equations and Dynamical Systems*. Springer, Berlin
55. Weinstein A (1973) Normal Modes for Nonlinear Hamiltonian Systems. *Inventiones Math* 20:47–57
56. Weinstein A (1978) Bifurcations and Hamilton Principle. *Math Z* 159:235–248
57. Whitehead GW (1995) *Elements of homotopy theory*, 3rd edn. Springer, Berlin
58. Wulf C, Roberts M (2002) Hamiltonian Systems Near Relative Periodic Orbits. *Siam J Appl Dyn Syst* 1(1):1–43
59. Yakubovich VA, Starzhinsky VM (1975) *Linear Differential Equations with Periodic coefficients*, vol 1:2. Wiley, UK

## Periodic Solutions of Non-autonomous Ordinary Differential Equations

JEAN MAWHIN

Département de Mathématique, Université Catholique de Louvain, Maryland, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Poincaré Operator and Linear Systems

Boundedness and Periodicity

Fixed Point Approach: Perturbation Theory

Fixed Point Approach: Large Nonlinearities

Guiding Functions

Lower and Upper Solutions

Direct Method of the Calculus of Variations

Critical Point Theory

Future Directions

Bibliography

### Glossary

**Banach fixed point theorem** If  $M$  is a complete metric space with distance  $d$ , and  $f: M \rightarrow M$  is contractive, i. e.  $d(f(u), f(v)) \leq \alpha d(u, v)$  for some  $\alpha \in [0, 1)$  and all  $u, v \in M$ , then  $f$  has a unique fixed point  $u^*$  and  $u^* = \lim_{k \rightarrow \infty} f^k(u_0)$  for any  $u_0 \in M$ .

**Brouwer degree** An integer  $d_B[f, \Omega]$  which ‘algebraically’ counts the number of zeros of any continuous mapping  $f: \overline{\Omega} \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$  such that  $0 \notin f(\partial\Omega)$ , and is invariant for sufficiently small perturbations of  $f$ . If  $f$  is of class  $C^1$  and its zeros are non degenerate, then  $d_B[f, \Omega] = \sum_{x \in f^{-1}(0)} \text{sign } \det f'(x)$ .

**Brouwer fixed point theorem** Any continuous mapping  $f: B \rightarrow B$ , with  $B$  is homeomorphic to the closed unit ball in  $\mathbb{R}^n$ , has at least one fixed point.

**Leray–Schauder degree** The extension  $d_{LS}[I - g, \Omega]$  of the Brouwer degree, where  $\Omega$  is an open bounded subset of the Banach space  $X$ , and  $g: \overline{\Omega} \rightarrow X$  is continuous,  $g(\overline{\Omega})$  is relatively compact and  $0 \notin (I - g)(\partial\Omega)$ .



**Leray–Schauder–Schafer fixed point theorem** If  $X$  is a Banach space,  $g: X \rightarrow X$  is a continuous mapping taking bounded subsets into relatively compact ones, and if the set of possible fixed points of  $\varepsilon g$  ( $\varepsilon \in [0, 1]$ ) is bounded independently of  $\varepsilon$ , then  $g$  has at least one fixed point.

**Ljusternik–Schnirelmann category** The Ljusternik–Schnirelmann category  $cat(M)$  of a metric space  $M$  into itself is the smallest integer  $k$  such that  $M$  can be covered by  $k$  sets contractible in  $M$ .

**Lower and upper solutions** A lower (resp. upper) solution of the periodic problem  $u'' = f(t, u)$ ,  $u(0) = u(T)$ ,  $u'(0) = u'(T)$  is a function  $\alpha$  (resp.  $\beta$ ) of class  $C^2$  such that  $\alpha''(t) \geq f(t, \alpha(t))$ ,  $\alpha(0) = \alpha(T)$ ,  $\alpha'(0) \geq \alpha'(T)$  (resp.  $\beta''(t) \leq f(t, \beta(t))$ ,  $\beta(0) = \beta(T)$ ,  $\beta'(0) \leq \beta'(T)$ ).

**Palais–Smale condition for a  $C^1$  function  $\varphi: X \rightarrow \mathbb{R}$**

Any sequence  $(u_k)_{k \in \mathbb{N}}$  such that  $(\varphi(u_k))_{k \in \mathbb{N}}$  is bounded and  $\lim_{k \rightarrow \infty} \varphi'(u_k) = 0$  contains a convergent subsequence.

**Poincaré operator** The mapping defined in  $\mathbb{R}^n$  by  $P_T: y \mapsto p(T; y)$ , where  $p(t; y)$  is the unique solution of the Cauchy problem  $x' = f(t, x)$ ,  $x(0) = y$ .

**Schauder fixed point theorem** If  $C$  is a closed bounded convex subset of a Banach space  $X$ , any continuous mapping  $g: C \rightarrow C$  such that  $g(C)$  is relatively compact has at least one fixed point.

**Sobolev inequality** For any function  $u \in L^2(0, T)$  such that  $u' \in L^2(0, T)$  and  $\int_0^T u(t) dt = 0$ , one has  $\max_{t \in [0, T]} |u(t)| \leq (T^{1/2}/2\sqrt{3}) [\int_0^T |u'(t)|^2 dt]^{1/2}$ .

**Wirtinger inequality** For any function  $u \in L^2(0, T)$  such that  $u' \in L^2(0, T)$  and  $\int_0^T u(t) dt = 0$ , one has  $\int_0^T |u(t)|^2 dt \leq (T^2/4\pi^2) \int_0^T |u'(t)|^2 dt$ .

## Definition of the Subject

Many phenomena in nature can be modeled by systems of ordinary differential equations which depend periodically upon time. For example, a linear or nonlinear oscillator can be forced by a periodic external force, and an important question is to know if the oscillator can exhibit a periodic response under this forcing. This question originated from problems in classical and celestial mechanics, before receiving important applications in radioelectricity and electronics. Nowadays, it also plays a great role in mathematical biology and population dynamics, as well as in mathematical economics, where the considered systems are often subject to seasonal variations. The general theory originated with Henri Poincaré's work in celestial

mechanics, at the end of the XIXth century, and has been constantly developed since.

## Introduction

To motivate the problem and its difficulties, let us start with the simple linear oscillator with forcing (or input)  $h \in L^2(0, 2\pi)$

$$L_\lambda u := -u'' - \lambda u = h(t), \quad (1)$$

where  $\lambda \in \mathbb{R}$ . We are interested in discussing the existence or non-existence, and the uniqueness or multiplicity of a  $2\pi$ -periodic solution  $u$  of (1). Let

$$h(t) \sim c_0 + \sum_{k=1}^{\infty} [c_k \cos kt + d_k \sin kt],$$

$$u(t) \sim a_0 + \sum_{k=1}^{\infty} [a_k \cos kt + b_k \sin kt],$$

with

$$c_0 = \frac{1}{2\pi} \int_0^{2\pi} h(t) dt,$$

$$\begin{Bmatrix} c_k \\ d_k \end{Bmatrix} = \frac{1}{\pi} \int_0^{2\pi} h(t) \begin{Bmatrix} \cos kt \\ \sin kt \end{Bmatrix} dt \quad (k = 1, 2, \dots)$$

and similarly for  $u$ , be the Fourier series of  $h$  and  $u$ . If  $u$  is a possible  $2\pi$ -periodic solution of (1),  $u$  is of class  $C^1$  and  $u'' \in L^2(0, 2\pi)$ , so that Fourier series of  $u$  and  $u'$  converge uniformly on  $[0, 2\pi]$  to  $u$  and  $u'$ . Furthermore, from Parseval equality

$$\|h\|_2^2 := \frac{1}{2\pi} \int_0^{2\pi} h^2(t) dt = c_0^2 + \frac{1}{2} \sum_{k=1}^{\infty} [c_k^2 + d_k^2],$$

and, as  $u'' \in L^2(0, 2\pi)$ ,  $u''(t) \sim -\sum_{k=1}^{\infty} k^2 [a_k \cos kt + b_k \sin kt]$ . Therefore, finding the  $2\pi$ -periodic solutions of (1) is equivalent to solving the infinite-dimensional linear system in  $l^2$  with unknowns  $(a_0, a_1, b_1, \dots)$

$$\begin{aligned} -\lambda a_0 &= c_0, & (k^2 - \lambda) a_k &= c_k, \\ (k^2 - \lambda) b_k &= d_k & (k = 1, 2, \dots). \end{aligned} \quad (2)$$

Letting

$$\Sigma := \{k^2: k = 0, 1, 2, \dots\},$$

we see that if  $\lambda \notin \Sigma$  (non-resonance), system (2) has the unique solution

$$a_0 = -\frac{c_0}{\lambda}, \quad a_k = \frac{c_k}{k^2 - \lambda}, \quad b_k = \frac{d_k}{k^2 - \lambda} \quad (k = 1, 2, \dots),$$

which gives the unique  $2\pi$ -periodic solution of (1)

$$\begin{aligned} u(t) &= (L_\lambda^{-1}h)(t) \\ &= -\frac{c_0}{\lambda} + \sum_{k=1}^{\infty} \frac{1}{k^2 - \lambda} [c_k \cos kt + d_k \sin kt]. \end{aligned}$$

Furthermore, using Parseval equality,

$$\|u\|_2^2 = \|L^{-1}h\|_2^2 = \frac{c_0^2}{\lambda^2} + \frac{1}{2} \sum_{k=1}^{\infty} \frac{1}{(k^2 - \lambda)^2} [c_k^2 + d_k^2] \quad (3)$$

$$\begin{aligned} &\leq \frac{1}{[\text{dist}(\lambda, \Sigma)]^2} \left[ c_0^2 + \frac{1}{2} \sum_{k=1}^{\infty} (c_k^2 + d_k^2) \right] \\ &= \frac{1}{[\text{dist}(\lambda, \Sigma)]^2} \|h\|_2^2. \end{aligned} \quad (4)$$

Now, if  $\lambda = j^2 \in \Sigma$  (resonance) and if  $c_j \neq 0$  or  $d_j \neq 0$ , then (2) has no solution and (1) has no  $2\pi$ -periodic solution. If  $c_j = d_j = 0$ , then (1) has the infinite family of  $2\pi$ -periodic solutions

$$u(t) = \alpha + \sum_{k=1}^{\infty} \frac{1}{k^2} [c_k \cos kt + d_k \sin kt] \quad (\alpha \in \mathbb{R})$$

if  $j = 0$  and

$$\begin{aligned} u(t) &= \alpha \cos jt + \beta \sin jt - \frac{c_0}{j^2} \\ &\quad + \sum_{\substack{k=1 \\ k \neq j}}^{\infty} \frac{1}{k^2 - j^2} [c_k \cos kt + d_k \sin kt] \end{aligned}$$

$(\alpha, \beta \in \mathbb{R})$  if  $j \neq 0$ .

For the nonlinear oscillator

$$-u'' = g(u) + h(t) \quad (5)$$

where  $g: \mathbb{R} \rightarrow \mathbb{R}$  is continuous, a nonlinear version of the non-resonant linear situation can be obtained. Indeed, assume that there exist numbers  $0 < \alpha \leq \beta$  such that, for all  $u \neq v \in \mathbb{R}$  one has

$$\alpha \leq \frac{g(u) - g(v)}{u - v} \leq \beta, \quad [\alpha, \beta] \cap \Sigma = \emptyset \quad (6)$$

which means that there exists  $j^2 \in \Sigma$  such that

$$j^2 < \alpha \leq \beta < (j+1)^2. \quad (7)$$

If  $\gamma := (\alpha + \beta)/2$ , so that  $\gamma \notin \Sigma$ , we can write (5) in the equivalent form

$$-u'' - \gamma u = g(u) - \gamma u + h(t), \quad (8)$$

and, if we define  $F_\gamma: L^2(0, 2\pi) \rightarrow L^2(0, 2\pi)$  by  $[F_\gamma(u)](t) = g(u) - \gamma u(t) + h(t)$ , it follows from (6) that

$$\|F_\gamma(u) - F_\gamma(v)\|_2 \leq \delta \|u - v\|_2, \quad (9)$$

with  $\delta = (\beta - \alpha)/2$ . Furthermore, finding the  $2\pi$ -periodic solutions of (8) is equivalent to solving the equation in  $L^2(0, 2\pi)$

$$u = L_\gamma^{-1} F_\gamma(u). \quad (10)$$

Now, using estimates (3), (7) and (9), we get, for all  $u, v \in L^2(0, 2\pi)$ ,

$$\|L_\gamma^{-1}[F_\gamma(u) - F_\gamma(v)]\|_2 \leq \frac{\delta}{\text{dist}(\gamma, \Sigma)} \|u - v\|_2,$$

with  $\delta/\text{dist}(\gamma, \Sigma) < 1$ . Banach fixed point theorem implies the existence of a unique fixed point  $u$  of  $L_\gamma^{-1}F_\gamma$ , and hence of a unique  $2\pi$ -periodic solution of (5).

Such a result, proved in 1949 by Dolph [14] for Dirichlet boundary conditions and in 1976 in [31] for the periodic problem, has been extended in various directions. For example, it was proved in [8], using sophisticated topological methods and delicate a priori estimates, that if  $G(u) = \int_0^u g(s)ds$ , the forced nonlinear oscillator (5) has at least one  $2\pi$ -periodic solution for each continuous  $h: [0, 2\pi] \rightarrow \mathbb{R}$  if  $g$  is odd,  $ug(u) \geq \theta G(u) > 0$  for some  $\theta \geq 1$  and all large  $|u|$ , and if

$$\left[ \liminf_{u \rightarrow +\infty} \frac{2G(u)}{u^2}, \limsup_{u \rightarrow +\infty} \frac{2G(u)}{u^2} \right] \neq \{k^2\}$$

for any positive integer  $k$ . However, it is still an open problem to know if the natural generalization of (6)

$$j^2 < \liminf_{|u| \rightarrow +\infty} \frac{2G(u)}{u^2} \leq \limsup_{|u| \rightarrow +\infty} \frac{2G(u)}{u^2} < (j+1)^2$$

is sufficient for the existence of a  $2\pi$ -periodic solution to (5). The problem of obtaining nonlinear versions of the resonance situation is more difficult and also requires more sophisticated tools. It will be considered in Sects. “Lower and Upper Solutions” to “Critical Point Theory”

## Poincaré Operator and Linear Systems

Let  $T > 0$  be fixed,  $f: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $(t, x) \mapsto f(t, x)$  be locally Lipschitzian in  $x$  and continuous. For each  $y \in \mathbb{R}^n$ , there exists a unique solution  $x(t) = p(t; y)$  of the Cauchy problem

$$x' = f(t, x), \quad x(0) = y,$$

defined and continuous on a maximal open set  $G = \{(t, y) \in \mathbb{R} \times \mathbb{R}^n : \tau_-(y) < t < \tau_+(y)\}$ , for some  $-\infty \leq \tau_-(y) < 0 < \tau_+(y) \leq +\infty$ .

A T-periodic solution of the differential system

$$x' = f(t, x) \quad (11)$$

is a solution of (11) defined at least over  $[0, T]$  and such that  $x(0) = x(T)$ . If we assume in addition that  $f(t, x) = f(t + T, x)$  for all  $t \in \mathbb{R}$  and  $x \in \mathbb{R}^n$ , a T-periodic solution of (11) can be continued as a solution defined over  $\mathbb{R}$  and such that  $x(t) = x(t + T)$  for all  $t \in \mathbb{R}$ . Poincaré already observed at the end of the XIXth century [43] that  $p(t; y)$  is a T-periodic solution of (11) if and only if  $y \in \mathbb{R}^n$  is such that  $\tau_+(y) > T$  and  $y = p(T; y)$ , i.e.  $y$  is a fixed point of Poincaré operator  $P_T$  defined by  $P_T(y) = p(T; y)$  for  $y \in \mathbb{R}^n$  such that  $\tau_+(y) > T$ .

A simple application is the case of a linear system

$$x' = A(t)x \quad (12)$$

where  $A: [0, T] \rightarrow \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$  is a continuous  $(n \times n)$ -matrix-valued function. Given  $y \in \mathbb{R}^n$ , the solution of (12) such that  $x(0) = y$  can be written  $x(t) = X(t)y$  for some  $n \times n$  matrix such that  $X(0) = I$ , the fundamental matrix of (12). The corresponding Poincaré operator, given by  $P_T(y) = X(T)y$ , has non-trivial fixed points if and only if  $I - X(T)$  is singular. If  $h: [0, T] \rightarrow \mathbb{R}^n$  is continuous, the solution of the forced linear system

$$x' = A(t)x + h(t) \quad (13)$$

such that  $x(0) = y$  being given by

$$p(t; y) = X(t)y + \int_0^t X(t)X^{-1}(s)h(s)ds, \quad (14)$$

the corresponding Poincaré operator is

$$P_T(y) = X(T)y + \int_0^T X(T)X^{-1}(s)h(s)ds. \quad (15)$$

Consequently, if  $I - X(T)$  is non singular, i.e. if (12) only has the trivial T-periodic solution  $x(t) \equiv 0$ , (15) has the unique fixed point

$$y = [I - X(T)]^{-1} \int_0^T X(T)X^{-1}(s)h(s)ds$$

and, by inserting this value of  $y$  in (14), (13) has the unique T-periodic solution

$$x(t) = \int_0^T G(t, s)h(s)ds, \quad (16)$$

where  $G(t, s)$  is the Green matrix explicitly given by

$$G(t, s) = \begin{cases} X(t)[I - X(T)]^{-1}X^{-1}(s) & \text{if } 0 \leq s \leq t \leq T \\ X(t)[I - X(T)]^{-1}X(T)X^{-1}(s) & \text{if } 0 \leq t < s \leq T \end{cases} \quad (17)$$

The situation is more complicated when  $I - X(T)$  is singular, which always happens in the simple case where  $A(t) \equiv 0$ , to which we restrict ourself. In this case,  $X(t) \equiv I$ , and  $P_T(y) = y + \int_0^T h(s)ds$ . It has fixed points if and only if  $h$  has mean value zero, namely

$$\bar{h} := \frac{1}{T} \int_0^T h(s)ds = 0,$$

in which case (13) with  $A(t) \equiv 0$  has the family of T-periodic solutions

$$x(t) = y + \int_0^t h(s)ds \quad (y \in \mathbb{R}^n).$$

### Boundedness and Periodicity

When  $n = 1$  and  $f: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  is T-periodic with respect to  $t$ , the local uniqueness assumption implies that two solutions  $p(t; y)$  and  $p(t; z)$  with  $y < z$  are such that  $p(t; y) < p(t; z)$  for all  $t$  where they are defined. Hence, if  $p(t; y)$  is defined for all  $t \geq 0$ ,  $p(t + nT; y)$  is the solution of (11) equal to  $y_n = p(nT; y)$  at  $t = 0$  for any integer  $n \geq 1$ . If  $y_1 = y$ ,  $p(t; y)$  is T-periodic; if, say,  $y_1 < y$ , then  $p(t + T; y) = p(t; y_1) < p(t; y)$  for all  $t \geq 0$ , and hence,  $p(t + (n + 1)T; y) < p(t + nT; y)$  for all  $t \geq 0$ . Thus, for any  $t \geq 0$ , the sequence  $(p(t + nT; y))_{n \in \mathbb{N}}$  is monotone. If  $p(t; y)$  is bounded in the future, i.e. if there exists  $M > 0$  such that  $|p(t; y)| \leq M$  for all  $t \geq 0$ , the sequence above, monotone, bounded and equicontinuous (as  $(p'(t + nT; y))_{n \in \mathbb{N}}$  is bounded), converges uniformly on each bounded interval to a continuous function  $\xi(t)$ . It follows from the identity

$$p(t + nT; y) = p(nT; y) + \int_0^t f(s, p(s + nT; y))ds$$

that  $\xi(t)$  is a solution of (11) defined for  $t \geq 0$  and that

$$\xi(T) = \lim_{n \rightarrow \infty} p((n + 1)T; y) = \lim_{n \rightarrow \infty} p(nT; y) = \xi(0).$$

This gives a result proved by Massera [29] in 1950: if  $n = 1$  and  $f$  is T-periodic in  $t$ , locally Lipschitzian in  $x$ , and continuous, then (11) admits a T-periodic solution if and only if it admits a solution bounded in the future. Of

course, the same statement holds if we replace ‘in the future’ by ‘in the past’. The result needs not to be true for  $n \geq 2$ , but, using delicate arguments of two-dimensional topology, Massera [29] has proved that if  $n = 2$ ,  $f$  is T-periodic in  $t$ , locally Lipschitzian in  $x$ , continuous, if all solutions of (11) exist in the future and if one of them is bounded in the future, then (11) admits a T-periodic solution. Again, the same statement holds if we replace ‘in the future’ by ‘in the past’. An important consequence of Massera’s results is that the absence of T-periodic solutions in a one or two-dimensional T-periodic system implies the unboundedness of all its solutions in the past and in the future.

By reinforcing Massera’s conditions, one can find criteria for the existence of T-periodic solutions valid for any  $n$ . A set  $G$  is a positively invariant set for (11) if, for each  $y \in G$ ,  $p(t; y) \in G$  for all  $t \in [0, \tau_+(y)]$ . In particular, if  $G$  is bounded, then  $\tau_+(y) = +\infty$ . Now, if  $G$  is invariant, and homeomorphic to the unit closed ball  $B[0; 1]$  in  $\mathbb{R}^n$ , Poincaré operator maps continuously  $G$  into itself, and Brouwer fixed point theorem implies that (11) has at least one T-periodic solution with values in  $G$ .

Such a result, which can be traced to Lefschetz [24] and to Levinson [25] in the early 1940s has been widely used to study the periodic solutions of the periodically forced Liénard differential equation

$$y'' + h(y)y' + g(y) = e(t), \quad (18)$$

written as a two-dimensional first order system, under various conditions upon the friction coefficient  $h(y)$ , the restoring force  $g(y)$  and the forcing term  $e(t)$ . Some of Levinson’s results have been at the origin of the theory of chaos.

### Fixed Point Approach: Perturbation Theory

Formula (16) suggests another approach for finding periodic solutions, which is independent of Cauchy’s problem and requires less regularity. Consider the nonlinear differential system

$$x' = A(t)x + f(t, x) \quad (19)$$

where  $A: [0, T] \rightarrow \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$  and  $f: [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  are continuous, and the corresponding linear system  $x' = A(t)x$  only has the trivial T-periodic solution. Using formula (16), finding the T-periodic solutions of (19) is equivalent to solving the nonlinear integral equation

$$x(t) = \int_0^T G(t, s)f(s, x(s))ds := [\mathcal{H}(x)](t) \quad (t \in [0, T]), \quad (20)$$

with  $G(t, s)$  defined in (17), in the Banach space  $C_T^\#$  of continuous functions such that  $x(0) = x(T)$ , i. e. to finding the fixed points of the nonlinear operator  $\mathcal{H}: C_T^\# \rightarrow C_T^\#$  defined in (20). Consider now the family of problems

$$x' = A(t)x + \varepsilon f(t, x), \quad x(0) = x(T) \quad (\varepsilon \in \mathbb{R}), \quad (21)$$

where  $x' = A(t)x$  only has the trivial T-periodic solution. Solving (21) is equivalent to solving the equation in  $C_T^\#$

$$\mathcal{K}(x, \varepsilon) := x - \varepsilon \mathcal{H}(x) = 0.$$

Trivially,  $\mathcal{K}(x, 0) = 0$  if and only if  $x = 0$ . If  $f'_x(t, x)$  exists and is continuous, the Fréchet derivative  $\mathcal{K}'_x(x, \varepsilon)$  at  $x \in C_T^\#$  is given by  $\mathcal{K}'_x(x, \varepsilon) = I - \varepsilon \mathcal{H}'(x)$ , so that  $\mathcal{K}'_x(0, 0) = I$  is invertible. It follows from the implicit function theorem in Banach spaces that for some  $\varepsilon_0 > 0$  and each  $|\varepsilon| \leq \varepsilon_0$ , (21) has a unique solution  $x_\varepsilon$  such that  $x_\varepsilon \rightarrow 0$  as  $\varepsilon \rightarrow 0$ . Such a result can be traced to Poincaré [43], who proved it using the associated Cauchy problem. From the proof of the implicit function theorem or from Banach fixed point theorem, it follows that  $x_\varepsilon$  can be obtained as the limit of the sequence of successive approximations given by

$$x_\varepsilon^0 = 0, \quad x_\varepsilon^{k+1}(t) = \int_0^T G(t, s)f(s, x_\varepsilon^k(s))ds \quad (k = 0, 1, 2, \dots).$$

The situation is more complicated when  $x' = A(t)x$  has nontrivial T-periodic solutions, and again we restrict ourselves to the important special case where  $A(t) \equiv 0$ , i. e. to the family of problems

$$x' = \varepsilon f(t, x), \quad x(0) = x(T) \quad (\varepsilon \in \mathbb{R}), \quad (22)$$

which admits, for  $\varepsilon = 0$ , the  $n$ -parameter family of T-periodic solutions  $x(t) = c$  ( $c \in \mathbb{R}^n$ ). To reduce (22) to a fixed point problem, one can first notice that, as easily verified, the linear problem

$$Lx := x' + x(0) = h(t), \quad x(0) = x(T)$$

has, for each continuous  $h: [0, T] \rightarrow \mathbb{R}^n$ , the unique solution

$$[L^{-1}h](t) = \bar{h} + \int_0^t [h(s) - \bar{h}]ds. \quad (23)$$

In particular,  $L^{-1}\bar{h} = \bar{h}$ . Now, for  $\varepsilon \neq 0$ , problem (22) is equivalent to

$$x' = (1 - \varepsilon)\overline{f(\cdot, x(\cdot))} + \varepsilon f(t, x), \quad x(0) = x(T), \quad (24)$$

which, using (23), is equivalent to the fixed point problem in  $C_T^\#$

$$\begin{aligned} x(t) &= x(0) + \overline{f(\cdot, x(\cdot))} \\ &\quad + \varepsilon \int_0^t [f(s, x(s)) - \overline{f(\cdot, x(\cdot))}] ds \\ &:= [\mathcal{M}(x, \varepsilon)](t) \end{aligned}$$

introduced by Mawhin in 1969 [30], i.e. to the equation in  $C_T^\#$

$$\mathcal{N}(x, \varepsilon) := x - \mathcal{M}(x, \varepsilon) = 0.$$

$\mathcal{N}(x, 0) = 0$  is equivalent to  $x(t) = x(0) + \overline{f(\cdot, x(\cdot))}$ , and hence to  $x(t) \equiv c$ , with  $c \in \mathbb{R}^n$  such that

$$F(c) := \frac{1}{T} \int_0^T f(s, c) ds = 0. \quad (25)$$

If  $c_0$  is a solution of (25) satisfying condition

$$\text{Jac } F(c_0) := \det \left[ \frac{1}{T} \int_0^T f'_x(s, c_0) ds \right] \neq 0, \quad (26)$$

the Fréchet derivative  $\mathcal{N}'_x(c_0, 0)$  is invertible, and the implicit function theorem in Banach spaces implies the existence of  $\varepsilon_0 > 0$  such that, for each  $|\varepsilon| \leq \varepsilon_0$ , equation  $\mathcal{N}(x, \varepsilon) = 0$  has a unique solution  $x_\varepsilon$  such that  $x_\varepsilon \rightarrow c_0$  as  $\varepsilon \rightarrow 0$ , and the same conclusion holds for (22).

This result is also a consequence of other perturbation methods for periodic solutions based upon Poincaré, averaging, Lyapunov–Schmidt or Cesari–Hale methods, and described in some of the books given in the Bibliography. The proof by iteration of the implicit function theorem implies that  $x_\varepsilon$  can be obtained as the limit of the sequence of successive approximations given by

$$\begin{aligned} x_\varepsilon^0 &= c_0, \\ x_\varepsilon^{k+1} &= c + \varepsilon [T - \mathcal{M}'_x(c_0, 0)]^{-1} \\ &\quad \cdot [\mathcal{M}(x_\varepsilon^k, \varepsilon) - \mathcal{M}(c_0, 0) - \mathcal{M}'_x(c_0, 0)(x_\varepsilon^k - c)] \\ &\quad (k = 0, 1, 2, \dots). \end{aligned}$$

For example, the search of positive periodic solutions of Verhulst equation with seasonal variations in population dynamics

$$y' = \varepsilon[a(t)y - b(t)y^2] \quad (\varepsilon > 0), \quad (27)$$

where  $a, b: \mathbb{R} \rightarrow (0, +\infty)$  are continuous and have period  $T$ , is equivalent, through the transformation  $y = e^u$ , to the problem

$$u' = \varepsilon[a(t) - b(t)e^u], \quad u(0) = u(T)$$

for which  $F(c) = \bar{a} - \bar{b}e^c$ . This equation has the unique solution  $c_0 = \log \bar{a}/\bar{b}$ , and  $F'(c_0) = -\bar{a} \neq 0$ . Hence, for sufficiently small  $\varepsilon > 0$ , Verhulst equation with seasonal variations (27) has at least one positive  $T$ -periodic solution  $y_\varepsilon$  such that  $y_\varepsilon \rightarrow \bar{a}/\bar{b}$  when  $\varepsilon \rightarrow 0$ .

### Fixed Point Approach: Large Nonlinearities

The use of more sophisticated fixed point theorems provides existence results which are not of perturbation type. If  $f$  is continuous,  $\mathcal{H}$  defined in (20) is continuous, and, using Arzelà–Ascoli theorem,  $\mathcal{H}$  takes bounded sets into relatively compact sets, i.e.  $\mathcal{H}$  is completely continuous on  $C_T^\#$ . Using Schauder fixed point theorem,  $\mathcal{H}$  has at least one fixed point if it maps a closed ball of  $C_T^\#$  into itself. It is the case in particular when  $f$  satisfies the growth condition

$$\|f(t, x)\| \leq \alpha \|x\| + \beta \quad ((t, x) \in [0, T] \times \mathbb{R}^n),$$

with  $\beta \geq 0$  and  $|\alpha| \int_0^T |G(t, s)| ds < 1$  for all  $t \in [0, T]$ , and, in particular, when  $f$  is bounded on  $[0, T] \times \mathbb{R}^n$ .

More general existence conditions can be deduced from Leray–Schauder–Schafer fixed point theorem, which, applied to  $\mathcal{H}$ , implies that (19) has at least one  $T$ -periodic solution if there exists  $R > 0$  such that any possible solution  $x$  of the family of problems

$$x'(t) = A(t)x + \varepsilon f(t, x), \quad x(0) = x(T) \quad (\varepsilon \in [0, 1])$$

is such that  $\max_{t \in [0, T]} \|x(t)\| < R$ . Special cases of this result, for particular equations, can be traced to Stoppelli [49], and the general form was first given by Reisig [46] and Villari [50].

For example, coming back to the equivalent form of the problem of positive periodic solutions of Verhulst equation with seasonal variations

$$u' = a(t) - b(t)e^u, \quad u(0) = u(T), \quad (28)$$

where  $a, b: \mathbb{R} \rightarrow (0, +\infty)$  are continuous and have period  $T$ , we associate to (28) the family of problems

$$u' + u = \varepsilon[a(t) + b(t)e^u + u], \quad u(0) = u(T), \quad (\varepsilon \in [0, 1]), \quad (29)$$

which reduces to (28) for  $\varepsilon = 1$ , and only has the trivial solution for  $\varepsilon = 0$ . If, for some  $\varepsilon \in [0, 1]$ , a possible solution  $u$  of (29) reaches a positive maximum at  $\tau$ , then

$$0 = u'(\tau) = \varepsilon[a(\tau) - b(\tau)e^{u(\tau)}] - (1 - \varepsilon)u(\tau),$$

so that  $\varepsilon[a(\tau) - b(\tau)e^{u(\tau)}] = (1 - \varepsilon)u(\tau) \geq 0$  and

$$u(\tau) \leq \log \frac{a(\tau)}{b(\tau)} \leq \log \frac{\max_{[0, T]} a}{\min_{[0, T]} b}.$$



Similarly, if  $u$  reaches a negative minimum at  $\tau'$ , then

$$u(\tau') \geq \log \frac{a(\tau')}{b(\tau')} \geq \log \frac{\min_{[0,T]} a}{\max_{[0,T]} b}.$$

The existence of at least one  $T$ -periodic solution  $u$  for (27) follows from the previous theorem with

$$R > \max \left( \log \frac{\max_{[0,T]} a}{\min_{[0,T]} b}, -\log \frac{\min_{[0,T]} a}{\max_{[0,T]} b} \right),$$

and it gives the positive solution  $y(t) = e^{u(t)}$  for the original Verhulst equation with seasonal variations  $y' = a(t)y - b(t)y^2$ , which reduces to the positive equilibrium  $y(t) \equiv \frac{a}{b}$  in the non-seasonal case where  $a$  and  $b$  are constant.

In the case of system (22), Schauder or Leray-Schauder-Schaefer fixed point theorems may be difficult to apply to  $\mathcal{M}(x, \varepsilon)$  because of the presence of the  $x(0)$  term. A generalization is required, based upon the concept of topological degree of some mappings  $\mathcal{F}$  defined on the closure  $\overline{\Omega}$  of a bounded open set  $\Omega$  of a Banach space  $X$ , and such that  $\mathcal{F}(x) \neq 0$  for  $x \in \partial\Omega$ . This topological degree, an integer counting ‘algebraically’ the number of zeros of  $\mathcal{F}$  in  $\Omega$ , is equal, for  $\mathcal{F} = \mathcal{I}$  to 1 or 0 according to  $0 \in \Omega$  or  $0 \notin \overline{\Omega}$ , implies the existence of a zero of  $\mathcal{F}$  in  $\Omega$  when it is nonzero, and the topological degree of  $\mathcal{F}(\cdot, \varepsilon)$  is independent of  $\varepsilon$  when  $\mathcal{F}(x, \varepsilon) \neq 0$  for all  $(x, \varepsilon) \in \partial\Omega \times [0, 1]$ .

Applied to the family of mappings  $\mathcal{I} - \mathcal{M}(x, \varepsilon)$  introduced in (22), Leray-Schauder and Brouwer degree theory implies the following continuation theorem, proved by Mawhin in 1969 [30]: if one can find an open bounded set  $\Omega \subset C_T^\#$  such that

- (i) for each  $\varepsilon \in (0, 1]$ , problem (22) has no solution on  $\partial\Omega$ ,
- (ii) system  $F(c) = 0$  defined in (25) has no solution on  $\partial\Omega \cap \mathbb{R}^n$ , with  $\mathbb{R}^n$  identified with constant functions in  $C_T^\#$ ,
- (iii) the Brouwer degree  $d_B[F, \Omega \cap \mathbb{R}^n]$  is different from zero,

then system (11) has at least one  $T$ -periodic solution in  $\Omega$ . If condition (i) is dropped, the conclusion remains valid for (22) with  $\varepsilon$  sufficiently small. Notice that in the conditions of the perturbation result described above, condition (ii) holds for  $\Omega$  a sufficient small open neighborhood of the zero  $c_0$  of  $F$ , and condition (26) implies that  $d_B[F, \Omega \cap \mathbb{R}^n] = \text{sign Jac } F(c_0) = \pm 1$ .

For example, consider the complex-valued Riccati-type equation

$$z' = p(t) + q(t)z + r(t)\widehat{z} + \widehat{z}^2 \quad (30)$$

where  $\widehat{z}$  denotes the complex conjugate of  $z$ , and  $p, q, r: [0, T] \rightarrow \mathbb{C}$  are continuous. It is nothing but a concise writing for a system of two real differential equations. If  $\varepsilon \in (0, 1]$  and  $z(t)$  is a possible  $T$ -periodic solution of

$$z' = \varepsilon[p(t) + q(t)z + r(t)\widehat{z} + \widehat{z}^2], \quad (31)$$

then, multiplying each member of (31) by  $z^2$  and integrating over  $[0, T]$ , we get

$$\begin{aligned} 0 &= \frac{1}{T} \int_0^T \left( \frac{z^3}{3} \right)' dt = \frac{1}{T} \int_0^T z^2(t) z'(t) dt \\ &= \varepsilon \left\{ \frac{1}{T} \int_0^T [p(t)z^2(t) + q(t)z^3(t) \right. \\ &\quad \left. + r(t)\widehat{z}(t)z^2(t) + |z(t)|^4] dt \right\}. \end{aligned}$$

Hence letting, for  $p \geq 1$ ,

$$\begin{aligned} \|u\|_p &= \left[ \frac{1}{T} \int_0^T |u(t)|^p dt \right]^{\frac{1}{p}}, \\ \|u\|_\infty &= \max_{t \in [0, T]} |u(t)|, \end{aligned}$$

and using Hölder inequality, we obtain

$$\|z\|_4^4 \leq \|p\|_2 \|z\|_4^2 + [\|q\|_4 + \|r\|_4] \|z\|_4^3$$

so that  $\|z\|_4 \leq R_1$ , where  $R_1$  is the positive root of equation  $r^2 = [\|q\|_4 + \|r\|_4]r + \|p\|_2$ . From (31) follows then that

$$\|z'\|_1 \leq \|p\|_1 + [\|q\|_{4/3} + \|r\|_{4/3}]R_1 + R_1^2 = R_2.$$

Inequalities upon  $\|z\|_4$  and  $\|z'\|_1$  imply that  $\|z\|_\infty < R$  for some  $R = R(R_1, R_2)$ . Now, for  $c \in \mathbb{C}$ ,

$$F(c) = \overline{p} + \overline{q}c + \overline{r}\widehat{c} + \widehat{c}^2$$

and hence, if  $F(c) = 0$ , we have

$$|c|^2 \leq (\|q\|_1 + \|r\|_1)|c| + \|p\|_1 \leq (\|q\|_4 + \|r\|_4)|c| + \|p\|_2$$

so that  $|c| \leq R_1 < R$ . Finally,  $d_B[F, B(0, R)] = -2$ . Hence all conditions of the second existence theorem hold for  $\Omega = B(0, R)$ , and (30) has at least one  $T$ -periodic solution, a result first proved, using a different approach, by Srzednicki [48] in 1994, the present proof being given in [35]. Notice that this result could not have been deduced from Leray-Schauder-Schaefer theorem, whose assumptions imply that the associated Leray-Schauder degree has absolute value one, although, for (30), its has absolute value two. The perturbed problem

$$z' = \varepsilon[p(t) + q(t)z + r(t)\overline{z} + z^2], \quad z(0) = z(T)$$

satisfies conditions (ii) and (iii) of Mawhin's continuation theorem for  $\Omega = B(0; R)$  with  $R$  sufficiently large, and has at least one  $T$ -periodic solution for  $|\varepsilon|$  sufficiently small. But Lloyd [28] and Campos-Ortega [6] have shown by examples that the problem

$$z' = p(t) + q(t)z + r(t)\bar{z} + z^2, \quad z(0) = z(T),$$

in contrast to (30), may have no  $T$ -periodic solution for some choice of  $p, q, r$ .

Finally, notice that the examples developed here illustrate two fundamental approaches to obtain a priori bounds for the possible  $T$ -periodic solutions: maximum principle-type argument and integral estimates.

### Guiding Functions

A useful variant of Mawhin's continuation theorem given in Sect. "Fixed Point Approach: Large Nonlinearities" has been proved in 1992 by Capietto, Mawhin and Zanolin [6], with another proof in [5] based upon degree theory for  $S^1$ -equivariant mappings. It states, for  $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$  and  $f: [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  continuous, that if there exists an open bounded set  $\Omega \subset C_T^\#$  such that the family of problems

$$x' = (1-\varepsilon)g(x) + \varepsilon f(t, x), \quad x(0) = x(T) \quad (\varepsilon \in [0, 1])$$

has no solution on  $\partial\Omega$ , and if  $d_B[g, \Omega \cap \mathbb{R}^n] \neq 0$ , then the problem

$$x' = f(t, x), \quad x(0) = x(T) \quad (32)$$

has at least one solution in  $\Omega$ . The proof is based upon the fact that  $|d_{LS}[I - \mathcal{M}, \Omega]| = |d_B[g, \Omega \cap \mathbb{R}^n]|$ , with  $\mathcal{M}$  the fixed point operator associated to the autonomous system  $x' = g(x)$ . Now a guiding function on  $G \subset \mathbb{R}^n$  for (32) is a function  $V: \mathbb{R}^n \rightarrow \mathbb{R}$  of class  $C^1$  such that, with  $(u|v)$  the inner product of  $u$  and  $v$  in  $\mathbb{R}^n$ ,

$$\|V'(x)\| \neq 0 \quad \text{and} \quad (V'(x)|f(t, x)) \leq 0 \\ \text{when} \quad (t, x) \in [0, 1] \times G.$$

This is an extension due to Mawhin-Ward [37] of a slight generalization given in [32] of a concept introduced in 1958 by Krasnosel'skii and Perov (see [22]) in the case where  $G = \mathbb{R}^n \setminus B(0; \rho)$  for some  $\rho > 0$ . For example  $V(x) = (1/2)\|x\|^2$  is a guiding function on  $\partial B(0; r)$  if  $(x|f(t, x)) \leq 0$  for  $(t, x) \in [0, T] \times \partial B(0, r)$ . We set  $V^r = \{x \in \mathbb{R}^n: V(x) < r\}$ . It is proved in [37] that if there exists a  $C^1$  function  $V: \mathbb{R}^n \rightarrow \mathbb{R}$  such that  $V^0$  is non-empty and bounded,  $V$  is a guiding function on  $V^{-1}(\{0\})$  for (32), and  $d_B[V', V^0] \neq 0$ , then problem (32)

has at least one solution with values in  $\overline{V^0}$ . The proof can be based upon the continuation theorem mentioned above with  $g(x) = -V'(x)$  and  $\Omega = \{x \in C_T^\#: x(t) \in G \ (t \in [0, T])\}$ , and its essential ingredient consists in showing that, for  $\lambda \in [0, 1)$ , the family of problems

$$x' = -(1-\varepsilon)V'(x) + \varepsilon f(t, x), \quad x(0) = x(T),$$

has no solution on  $\partial\Omega$ , which is done by studying the maximum of  $V(x(t))$ . In particular, if one takes  $G = \mathbb{R}^n \setminus B(0; \rho)$  for some  $\rho > 0$ , and  $V$  coercive, i. e.  $\lim_{\|x\|} V(x) = +\infty$ , one can show that  $d_B[V', V^0] = 1$ , and the existence follows. This generalizes the result about positively invariant sets mentioned in Sect. "Boundedness and Periodicity".

The following generalization of the concept of guiding function is also given in [37]. An averaged guiding function on  $G \subset \mathbb{R}^n$  for (32) is a function  $V: G \rightarrow \mathbb{R}$  of class  $C^1$  such that

$$\|V'(x)\| \neq 0 \quad \text{and} \quad \int_0^T (V'(x(t))|f(t, x(t)))dt \leq 0$$

when  $x \in C_T^\#$  and  $x(t) \in G$  for all  $t \in [0, T]$ . It is proved in [37] that problem (32) has at least a solution if there exists a  $C^1$  function  $V: \mathbb{R}^n \rightarrow \mathbb{R}$  such that  $V^0$  is non-empty,  $V^r$  is bounded for

$$r = \max \left\{ T \max_{u \in V^{-1}(\{0\})} \|V'(u)\|^2, \right. \\ \left. \max_{u \in V^{-1}(\{0\})} \int_0^T |(V'(u)|f(t, u))|dt \right\},$$

$V$  is an averaged guiding function on  $\mathbb{R}^n \setminus V^0$  for (32) and  $d_B[V', V^0] \neq 0$ .

Of course, any guiding function is an averaged guiding function, but the converse is false. On the other hand, any  $V$  such that  $(V'(x)|f(t, x)) \leq \alpha(t)$  for some nonnegative  $\alpha \in L^1(0, T)$ , all  $(t, x) \in [0, T] \times \mathbb{R}^n$ , and such that

$$\int_0^T \limsup_{\|x\| \rightarrow \infty} (V'(x)|f(t, x))dt < 0$$

is an averaged guiding function. In particular, taking  $V(x) = \|x\| - \log(1 + \|x\|)$ , so that  $V'(x) = \frac{x}{1+\|x\|}$ , it is easy to see that (32) has at least one solution if  $\left(\frac{x}{\|x\|}|f(t, x)\right) \leq \alpha(t)$  and

$$\int_0^T \limsup_{\|x\| \rightarrow \infty} \left(\frac{x}{\|x\|}|f(t, x)\right)dt < 0.$$

## Lower and Upper Solutions

A powerful way of proving the existence of  $T$ -periodic solutions for first or second order scalar differential equations is the *method of lower and upper solutions* (or *sub- and supersolutions*), introduced by Scorza Dragoni [47] in 1931 for Dirichlet boundary conditions and by Knobloch [21] in 1967 for periodic solutions. We describe it here, following the approach initiated in [33], for problem

$$u'' = f(t, u), \quad u(0) = u(T), \quad u'(0) = u'(T), \quad (33)$$

where  $f: [0, T] \times \mathbb{R} \rightarrow \mathbb{R}$  is continuous. The function  $\alpha$  (resp.  $\beta$ )  $\in C^2([0, T])$  is a *lower* (resp. *upper*) solution for (33) if

$$\begin{aligned} \alpha''(t) &\geq f(t, \alpha(t)) \quad (t \in ]0, T[), \\ \alpha(0) &= \alpha(T), \quad \alpha'(0) \geq \alpha'(T) \\ (\text{resp. } \beta''(t) &\leq f(t, \beta(t)) \quad (t \in ]0, T[), \\ \beta(0) &= \beta(T), \quad \beta'(0) \leq \beta'(T)). \end{aligned}$$

The method of lower and upper solutions reduces the existence of a  $T$ -periodic solution to that of an ordered couple of lower and upper solutions: if there exists a lower solution  $\alpha$  and an upper solution  $\beta$  for (33), such that  $\alpha(t) \leq \beta(t)$  for all  $t \in [0, T]$ , then (33) has at least one solution  $u$  such that  $\alpha(t) \leq u(t) \leq \beta(t)$  for all  $t \in [0, T]$ .

To prove such a result, one first introduce a modified problem whose solutions are solutions of (33), namely

$$\begin{aligned} u'' - u &= f(t, \gamma(t, u)) - \gamma(t, u), \\ u(0) &= u(T), \quad u'(0) = u'(T), \end{aligned} \quad (34)$$

where  $\gamma(t, u) = \alpha(t)$ ,  $u$  or  $\beta(t)$  according to  $u < \alpha(t)$ ,  $\alpha(t) \leq u \leq \beta(t)$  or  $u > \beta(t)$ . Notice that (33) and (34) coincide when  $\alpha(t) \leq u \leq \beta(t)$ , that its linear part only has the trivial  $T$ -periodic solution, and that its right-hand member is continuous and bounded everywhere. A result in Sect. “Fixed Point Approach: Large Nonlinearities” implies that (34) has at least one solution, say  $\tilde{u}$ . A contradiction argument, based upon simple characterizations of a maximum or a minimum, implies that  $\alpha(t) \leq \tilde{u}(t) \leq \beta(t)$  for all  $t \in [0, T]$ , so that  $\tilde{u}$  is a solution of (33).

A simple consequence is an elegant necessary and sufficient existence condition first proved by Kazdan and Warner [20] in 1975 in the Dirichlet case: if  $f(t, c)$  is non-decreasing in  $c$  for each fixed  $t$ , problem (33) has at least one solution if and only if  $\frac{1}{T} \int_0^T f(t, c) dt = 0$  for some  $c \in \mathbb{R}$ . Indeed, if (33) has a solution  $u$ , with  $u_m := \min_{[0, T]} u$ ,  $u_M := \max_{[0, T]} u$ , integrating both members of (33) over

$[0, T]$  and using the monotonicity of  $f(t, \cdot)$  gives

$$\int_0^T f(t, u_m) dt \leq 0 \leq \int_0^T f(t, u_M) dt,$$

and the necessity follows. For the sufficiency, if  $\int_0^T f(t, c) dt = 0$ , and  $v(t)$  is the unique solution of the linear problem

$$v'' = f(t, c), \quad v(0) = 0 = v(T), \quad v'(0) = v'(T),$$

then  $\alpha(t) := c - \max_{[0, T]} v + v(t) \leq c \leq \beta(t) := c - \min_{[0, T]} v + v(t)$  are ordered lower and upper solutions for (33).

In particular, the problem

$$u'' + g(u) = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

with  $g: \mathbb{R} \rightarrow \mathbb{R}$  continuous non increasing and  $h: [0, T] \rightarrow \mathbb{R}$  continuous, has at least one solution if and only if  $\bar{h} \in g(\mathbb{R})$ . This is a first nonlinear generalization of resonance at the first eigenvalue zero. For example, the problem

$$u'' - \arctan u = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

has at least one solution if and only if  $-\frac{\pi}{2} < \bar{h} < \frac{\pi}{2}$ , and the problem

$$u'' - u^+ = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

has at least one solution if and only if  $\bar{h} \leq 0$ . This last problem is an example of differential equation with asymmetric or jumping nonlinearities. Since the pioneering work of Fučík [18] in 1978, the study of problems of the type

$$\begin{aligned} u'' + \alpha u^+ - \beta u^- + g(u) &= h(t), \\ u(0) &= u(T), \quad u'(0) = u'(T) \end{aligned}$$

where the linear part in the nonlinear oscillator is replaced by a piecewise linear one, has been intensively studied. See for example [17] for references.

Combined with degree arguments, the method of lower and upper solutions also allows to obtain multiplicity results for  $T$ -periodic solutions. For example, Fabry, Mawhin and Nkashama [16] have proved in 1986 for the problem (with  $c \in \mathbb{R}$ ,  $s \in \mathbb{R}$ )

$$\begin{aligned} u'' + cu' + g(u) &= h(t) + s, \\ u(0) &= u(T), \quad u'(0) = u'(T) \end{aligned} \quad (35)$$

with  $\lim_{|u| \rightarrow \infty} g(u) = +\infty$ , the existence of  $s_0 \in \mathbb{R}$  such that problem (35) has no solution for  $s < s_0$ , at least

one solution for  $s = s_0$  and at least two solutions for  $s > s_0$ . Such a result is generally referred to as an Ambrosetti–Prodi problem, after the pioneering work of Ambrosetti and Prodi [4], in 1969, for elliptic boundary value problems. Ortega has studied the stability of the  $T$ -periodic solutions when  $c > 0$  (see [40]).

The example of

$$u'' + u = \sin t, \quad u(0) = u(2\pi), \quad u'(0) = u'(2\pi)$$

which has no solution (resonance at the second eigenvalue 1) and has the (unordered) lower and upper solution  $\alpha \equiv 1$  and  $\beta \equiv -1$  shows that the method of lower and upper solutions fails in general in the case of unordered lower and upper solutions. However, Amann, Ambrosetti and Mancini [2] have proved in 1978, for some elliptic boundary value problems, that existence may still hold under further conditions, using the following approach. For each  $\tilde{h} \in \tilde{C}_T^\# = \{h \in C_T^\# : \bar{h} = 0\}$ , the linear problem

$$\tilde{L}u := u'' = \tilde{h}(t), \quad u(0) = 0 = u(T), \quad u'(0) = u'(T)$$

has the unique solution

$$u(t) = [\tilde{L}^{-1}h](t) = \int_0^T K(t, s)\tilde{h}(s)ds,$$

where  $K(t, s) = -s/T(T-t)$  if  $0 \leq s \leq t \leq T$  and  $K(t, s) = -t/T(T-s)$  if  $0 \leq t < s \leq T$ . Hence  $u$  is a  $T$ -periodic solution of (33) if and only if  $u \in C_T^\#$  is a solution of

$$u(t) - u(0) = \int_0^T K(t, s)[f(s, u(s)) - \overline{f(\cdot, u(\cdot))}]ds \quad (36)$$

$$\overline{f(\cdot, u(\cdot))} = 0. \quad (37)$$

Letting  $c = u(0)$ ,  $y(t) = u(t) - u(0)$ , so that  $y \in \widehat{C}_T^\# = \{x \in C_T^\# : x(0) = 0\}$ , the first equation in (36) becomes

$$\begin{aligned} y(t) &= \int_0^T K(t, s)[f(s, c + y(s)) - \overline{f(\cdot, c + y(\cdot))}]ds \\ &:= [\mathcal{R}(y, c)](t), \end{aligned} \quad (38)$$

with  $\mathcal{R}$  is completely continuous in  $\widehat{C}_T^\# \times \mathbb{R}$ . Assuming now in addition that  $|f|$  is bounded on  $[0, T] \times \mathbb{R}$ , say by  $M$ , Leray–Schauder degree theory applied to (38) as a fixed point problem in  $y$  with  $c$  as a parameter implies that the set  $(y, c) \in \widehat{C}_T^\# \times \mathbb{R}$  satisfying (38) contains a continuum  $C$  whose projection on  $\widehat{C}_T^\#$  is contained in a ball  $B[0; R]$  with

$R$  depending only upon  $T$  and  $M$ , and whose projection on  $\mathbb{R}$  is  $\mathbb{R}$ . On  $C$ , we have, differentiating (38),

$$\begin{aligned} y'' &= f(t, c + y) - \overline{f(\cdot, c + y(\cdot))}, \\ y(0) &= y(T), \quad y'(0) = y'(T). \end{aligned} \quad (39)$$

Hence, if there exists  $(c, y) \in C$  such that  $\overline{f(\cdot, c + y(\cdot))} = 0$ , the second equation in (36) is also satisfied and  $c + y$  is a solution of (33). If not, by connectivity, either  $\overline{f(\cdot, c + y(\cdot))} < 0$  or  $\overline{f(\cdot, c + y(\cdot))} > 0$  for all  $(c, y) \in C$ . Assume now that (33) has a lower solution  $\alpha$  and an upper solution  $\beta$  which are not ordered, and consider, say the case where  $\overline{f(\cdot, c + y(\cdot))} < 0$  on  $C$ , the other one being similar. For each  $c \in \mathbb{R}$ , it follows from (39) that  $c + y$  is a lower solution for (33) whenever  $(c, y) \in C$ . Taking  $c$  such that  $c + y(t) \leq \beta(t)$  for all  $t \in [0, T]$  gives a couple of ordered lower and upper solution, and hence a  $T$ -periodic solution of (33). Proceeding like in the ordered case, we deduce from this result that if  $f$  is bounded and  $f(t, \cdot)$  non-increasing for each fixed  $t \in [0, T]$ , then (33) has at least one solution if and only if  $1/T \int_0^T f(t, c)dt = 0$  for some  $c \in \mathbb{R}$ . In particular, the problem

$$u'' + g(u) = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

with  $g: \mathbb{R} \rightarrow \mathbb{R}$  continuous, bounded and non decreasing, and  $h: [0, T] \rightarrow \mathbb{R}$  continuous, has at least one solution if and only if  $\bar{h} \in g(\mathbb{R})$ . This is another nonlinear generalization of resonance at the first eigenvalue 0. For example, the problem

$$u'' + \arctan u = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

as at least one solution if and only if  $-\frac{\pi}{2} < \bar{h} < \frac{\pi}{2}$ , and the problem

$$u'' + \frac{u^+}{1 + |u|} = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

has at least one solution if and only if  $0 \leq \bar{h} < 1$ . Notice that the boundedness condition upon  $f$  can be replaced by suitable linear growth conditions.

The case of constant lower and upper solutions gives a simple but useful existence condition: if  $f(t, \alpha) \leq 0 \leq f(t, \beta)$  for some  $\alpha \leq \beta$  and all  $t \in [0, T]$ , problem (33) has at least one solution  $u$  with  $\alpha \leq u(t) \leq \beta$  for all  $t \in [0, T]$ . The same conclusion holds for  $f$  bounded and  $f(t, \beta) \leq 0 \leq f(t, \alpha)$  for all  $t \in [0, T]$ . For example, taking  $\beta = R = -\alpha$  for sufficiently large  $R > 0$ , the problem

$$u'' = p(u) + h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

has at least one solution for each continuous  $h$ , when  $p$  is a real polynomial of odd order whose highest order term has a positive coefficient. This result can be traced to Lichtenstein [26], who proved it in 1915 using a variational method, and can be applied to the original Duffing's equation (with  $a > 0$ )

$$u'' + a \left( u - \frac{u^3}{6} \right) = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T)$$

introduced in 1918 by Duffing [15] as a nonlinear approximation to the forced pendulum equation.

However, for the exact pendulum equation

$$u'' + a \sin u = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T) \quad (40)$$

a necessary condition for the existence of a solution is  $-a \leq \bar{h} \leq a$ , as follows from integrating both members over  $[0, T]$ . To obtain a necessary and sufficient existence condition for (40), Dancer [12] has used in 1982 an approach similar to the one described for the case of unordered lower and upper solutions. Writing  $h = \bar{h} + \tilde{h}$ , he has proved, for each  $\tilde{h}$ , the existence of a (possibly degenerate) closed interval  $\mathcal{I}_{\tilde{h}} \subset [-a, a]$  such that (40) has at least one solution if and only if  $\bar{h} \in \mathcal{I}_{\tilde{h}}$ . In 1984, using degree theory, Mawhin and Willem [38] have proved the existence of at least two solutions for (40) when  $\bar{h} \in \text{int } \mathcal{I}$ . The same authors have proved that the set of  $\tilde{h}$  for which  $\mathcal{I}_{\tilde{h}}$  has a non-empty interior is open and dense in the space of continuous functions with mean value zero, but it is still an open problem to know if there exists or not some  $\tilde{h}$  such that  $\mathcal{I}_{\tilde{h}}$  is a singleton. See [36] for a survey of the various results related to the forced pendulum equation.

## Direct Method of the Calculus of Variations

When a differential equation or system can be written as the Euler-Lagrange equations of a problem of the calculus of variations, the direct method can be used to prove the existence of periodic solutions. Its fundamental result is that if  $X$  is a reflexive Banach space and if a weakly lower semi-continuous (w.l.s.c.) function  $\varphi: X \rightarrow (-\infty, +\infty]$  has a bounded minimizing sequence, then it has a minimum on  $X$ . Recall that  $\varphi$  is weakly lower semi-continuous at  $a \in X$  if  $\liminf_{k \rightarrow \infty} \varphi(u_k) \geq \varphi(a)$  whenever  $u_k \rightharpoonup a$ , and that  $(u_k)_{k \in \mathbb{N}}$  is a minimizing sequence for  $\varphi$  if  $\varphi(u_k) \rightarrow \inf_X \varphi$ . Another result, valid for an arbitrary Banach space  $X$ , is that a function  $\varphi: X \rightarrow \mathbb{R}$  of class  $C^1$ ,

bounded from below and satisfying the Palais-Smale condition has a minimum on  $X$ . The Palais-Smale condition was introduced by Palais and Smale [42] in 1964.

For simplicity, only the case of Lagrangian systems of differential equations of the type

$$u'' = F'_u(t, u) + \tilde{h}(t), \quad u(0) = u(T), \quad u'(0) = u'(T) \quad (41)$$

will be considered, where  $F: [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}$  is continuous together with  $F'_u: [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $\tilde{h}: [0, T] \rightarrow \mathbb{R}^n$  is continuous with  $\int_0^T \tilde{h}(t) dt = 0$ . We denote by  $H_T^1$  the Sobolev space

$$\{u \in L^2([0, T], \mathbb{R}^n): u \text{ has a weak derivative} \\ u' \in L^2([0, T], \mathbb{R}^n), u(0) = u(T)\},$$

with the inner product and norm

$$\langle u, v \rangle = \int_0^T [(u(t), v(t)) + (u'(t), v'(t))] dt, \\ \|u\|_{1,1} = \langle u, u \rangle^{\frac{1}{2}}.$$

It is easy to show that the action functional associated to (41) given by

$$\varphi(u) := \int_0^T \left[ \frac{\|u'(t)\|^2}{2} + F(t, u(t)) + (\tilde{h}(t), u(t)) \right] dt \quad (42)$$

$$= \int_0^T \left[ \frac{\|u'(t)\|^2}{2} + F(t, u(t)) + (\tilde{h}(t), \tilde{u}(t)) \right] dt \quad (43)$$

is well defined, w.l.s.c. and of class  $C^1$  on  $H_T^1$ , with

$$\varphi'(u)[v] \\ = \int_0^T [(u'(t), v'(t)) + (F'_u(t, u(t)) + \tilde{h}(t), v(t))] dt$$

for all  $u, v \in H_T^1$ . Furthermore, if  $u$  is a critical point of  $\varphi$ , i. e. if  $\varphi'(u) = 0$ , then  $u \in C^2([0, T], \mathbb{R}^n)$  and satisfies (41).

As a first application, following Mawhin-Willem [39], assume that  $\|F'_u\|$  is bounded, say by  $M$ , on  $[0, T] \times \mathbb{R}^n$  and that  $F$  satisfied the condition

$$\lim_{\|u\| \rightarrow \infty} \frac{1}{T} \int_0^T F(t, u) dt = +\infty, \quad (44)$$

first introduced by Ahmad-Lazer-Paul [1] in 1976 for elliptic boundary value problems. Writing  $u = \bar{u} + \tilde{u}$ , and



using Wirtinger's inequality, we obtain

$$\begin{aligned}\varphi(u) &= \int_0^T \frac{\|u'(t)\|^2}{2} dt + \int_0^T F(t, \bar{u}) dt \\ &\quad + \int_0^T \int_0^1 (F'_u(t, \bar{u} + s\tilde{u}(t)) + \tilde{h}(t), \tilde{u}(t)) dt \\ &\geq \int_0^T \frac{\|u'(t)\|^2}{2} dt - \frac{M'T^{\frac{3}{2}}}{2\pi} \left[ \int_0^T \frac{\|u'(t)\|^2}{2} dt \right]^{\frac{1}{2}} \\ &\quad + \int_0^T F(t, \bar{u}) dt\end{aligned}$$

with  $M' = M + \|\tilde{h}\|_\infty$ , which implies that  $\varphi(u) \rightarrow +\infty$  as  $\|u\|_{1,1} \rightarrow \infty$ . Hence all minimizing sequences for  $\varphi$  are bounded and the minimum of  $\varphi$  is reached at some  $u \in H_T^1$ , and is a solution of (41).

For example, consider the problem

$$u'' + g(u) = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T),$$

where  $h: [0, T] \rightarrow \mathbb{R}$  is continuous,  $g: \mathbb{R} \rightarrow \mathbb{R}$  is continuous, bounded and, with  $G(u) = \int_0^u g(s) ds$ , is such that

$$\widehat{g}(\pm\infty) := \lim_{u \rightarrow \pm\infty} \frac{G(u)}{u} \quad (45)$$

exist. For this problem

$$\lim_{|u| \rightarrow \infty} \int_0^T F(t, u) dt = \lim_{|u| \rightarrow \infty} u \left[ \bar{h} - \frac{G(u)}{u} \right] = +\infty$$

if  $\widehat{g}(+\infty) < \bar{h} < \widehat{g}(-\infty)$ . Such a condition, first introduced by Alonso and Ortega [2] in 1996, generalizes the one introduced by Lazer and Leach [23] in 1969 (with  $g(u)$  instead of  $G(u)/u$  in (45)), but usually referred as a Landesman–Lazer condition. For example the problem

$$\begin{aligned}u'' - \arctan u + a \sin u &= h(t), \\ u(0) &= u(T), \quad u'(0) = u'(T)\end{aligned}$$

with  $a \in \mathbb{R}$  has at least one solution if  $-\frac{\pi}{2} < \bar{h} < \frac{\pi}{2}$ .

As a second application, due to Willem [51], consider the case of a spatially periodic  $F$ , i. e. such that

$$F(t, u + T_i e_i) = F(t, u) \quad (1 \leq i \leq n) \quad (46)$$

for some  $T_i > 0$  and all  $(t, u) \in [0, T] \times \mathbb{R}^n$ , where the  $e_i$  denote the unit vectors in  $\mathbb{R}^n$  ( $1 \leq i \leq n$ ). As  $F$  is bounded over  $[0, T] \times \mathbb{R}^n$ , it is easy to see, using Sobolev inequality, that

$$\begin{aligned}\varphi(u) &\geq \frac{1}{2} \int_0^T \|u'(t)\|^2 dt \\ &\quad - C_1 \left[ \int_0^T \|u'(t)\|^2 dt \right]^{\frac{1}{2}} - C_2\end{aligned}$$

for some  $C_1, C_2 \geq 0$ , which implies the existence of  $C_3 > 0$  such that any minimizing sequence  $(u_k)_{k \in \mathbb{N}}$  for  $\varphi$  satisfies  $\int_0^T \|u_k(t)\|^2 dt \leq C_3$ . On the other hand, it follows from (46) that  $\varphi$  is such that

$$\varphi(u + T_i e_i) = \varphi(u) \quad (u \in H_T^1, 1 \leq i \leq n) \quad (47)$$

and hence  $(u_k + v_k)_{k \in \mathbb{N}}$  with  $v_{k,i} = k_i T_i$ , ( $k_i \in \mathbb{Z}$ ,  $1 \leq i \leq n$ ) is also a minimizing sequence. So there is always a minimizing sequence  $u_k$  such that  $0 \leq \bar{u}_{k,i} \leq T_i$  ( $1 \leq i \leq n$ ), and hence a bounded minimizing sequence in  $H_T^1$ , implying the existence of a  $T$ -periodic solution of (41) when (46) holds. In particular, this result implies that the periodic problem for the forced pendulum equation (40) has at least one solution for each  $a \in \mathbb{R}$  and each  $h$  with  $\bar{h} = 0$ , so that  $0 \in \mathcal{T}_h$ . Such a result, already proved by Hamel [19] in 1922 using calculus of variations, was independently rediscovered by Willem [51] in 1981 and Dancer [12] in 1982. Notice that, for  $n = 1$ , Dancer and Ortega [13] have proved in 2004 that if the minimum is isolated as a critical point of the action functional, then it is unstable in the sense of Lyapunov.

In this case of a periodic potential, multiplicity results can be proved using more sophisticated tools of the calculus of variations. Property (47) shows that one can consider naturally  $\varphi$  on the manifold  $\mathbb{T}^n \times \widetilde{H}_T^1$ , where  $\mathbb{T}^n$  denotes the  $n$ -torus and  $\widetilde{H}_T^1$  the subspace of  $H_T^1$  of functions with mean value zero. In such a case, a result of Palais [41] implies that if  $\varphi$  is bounded from below and satisfies the Palais–Smale condition, then  $\varphi$  has at least  $\text{cat}(\mathbb{T}^n \times \widetilde{H}_T^1)$  critical points. In this statement,  $\text{cat}(M)$  denotes the Lusternik–Schnirelmann category of the set  $M$  in itself, i. e. the least integer  $k$  such that  $M$  can be covered by  $k$  contractible subsets in  $M$ , a concept introduced by Ljusternik and Schnirelmann in 1934 [27]. Now it can be shown that

$$\text{cat}(\mathbb{T}^n \times \widetilde{H}_T^1) = \text{cat}(\mathbb{T}^n) = n + 1,$$

and hence, under condition (46), system (41) has at least  $n + 1$  geometrically distinct  $T$ -periodic solutions. In particular, when  $\bar{h} = 0$ , the forced pendulum equation (40) has at least two  $T$ -periodic solutions, a result first obtained in 1984 by Mawhin and Willem [38], using another variational approach. For other results in the spirit of Lusternik–Schnirelmann category, see [9,34,45]. Using, instead of Ljusternik–Schnirelmann category, another variational technique, namely Morse theory (see [10]), one can prove the existence of at least  $2^n$  geometrically distinct  $T$ -periodic solutions when they are non-degenerate.

In the case of a spatially periodic Hamiltonian system

$$Ju' = H'_u(t, u), \quad u(0) = u(T),$$

where  $H: [0, T] \times \mathbb{R}^{2n} \rightarrow \mathbb{R}$ ,  $H'_u: [0, T] \times \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$  are continuous, and  $J = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix}$  is the symplectic matrix, a similar result, namely the existence of at least  $2n + 1$  T-periodic solutions, has been proved by Conley and Zehnder [11] in 1984 using some finite-dimensional reduction and Conley's index. It solves a conjecture of Arnold in symplectic geometry (see e.g. [10]).

### Critical Point Theory

A function  $\varphi$  unbounded from below and from above needs not to have a minimum or a maximum, but may have a critical point of saddle point type. Many existence theorems for critical points of saddle point type of some functionals defined on a Banach space or a Banach manifold have been developed in the last thirty years, and have found many applications in the study of periodic solutions of differential systems having a variational structure, like Lagrangian or Hamiltonian systems. One such result, both simple and useful, is the saddle point theorem proved in 1978 by Rabinowitz [44]. Let  $\varphi: X \rightarrow \mathbb{R}$  be of class  $C^1$  on a Banach space  $X$  which can be splitted as a direct sum of closed subspaces  $X^- \oplus X^+$ , with  $X^+$  finite-dimensional. Assume that, for some  $R > 0$ ,  $\sup_{\partial B[0, R] \cap X^-} \varphi < \inf_{X^+} \varphi$ , and let

$$M := \{g: B[0, R] \cap X^- \rightarrow X^-, \text{ continuous, } g(s) = s \text{ on } \partial B[0, R] \cap X^-\}.$$

If  $\varphi$  satisfies a Palais–Smale condition on  $X$ ,  $\varphi$  has at least one critical point  $u$  such that  $\varphi(u) = c = \inf_{g \in M} \sup_{s \in B[0, R] \cap X^-} \varphi(g(s))$ .

As a simple application, we return to problem (41) with  $F'_u$  bounded over  $[0, T] \times \mathbb{R}^n$ , and we now assume instead of (44) that

$$\lim_{\|u\| \rightarrow \infty} \int_0^T F(t, u) dt = -\infty. \quad (48)$$

It is easy to see that  $\varphi$  is unbounded from below on the subspace of constant functions and bounded from below but unbounded from above on the subspace of functions with mean value zero. Hence, in particular,  $\inf_{\tilde{H}_T^1} \varphi > -\infty$ , and, for all sufficiently large  $R > 0$ ,  $\sup_{\partial B[0, R] \cap \tilde{H}_T^1} \varphi < \inf_{\tilde{H}_T^1} \varphi$ , where  $\tilde{H}_T^1$  is the subspace of constant functions, so that  $H_T^1 = \tilde{H}_T^1 \oplus \tilde{H}_T^1$ . The verification of the Palais–Smale condition follows from getting first a bound on  $\tilde{u}_k$  using the boundedness of  $F'_u$  and then a bound for  $\tilde{u}_k$  using condition (48). Hence (41) has at least one T-periodic solution. In particular, the problem

$$u'' + g(u) = h(t), \quad u(0) = u(T), \quad u'(0) = u'(T),$$

where  $h: [0, T] \rightarrow \mathbb{R}$  is continuous and  $g: \mathbb{R} \rightarrow \mathbb{R}$  is continuous, bounded and such that (with  $\widehat{g}(\pm\infty)$  defined in (45))  $\widehat{g}(-\infty) < \bar{h} < \widehat{g}(+\infty)$ , has at least one solution. For example, problem

$$u'' + \arctan u + a \sin u = h(t), \\ u(0) = u(T), \quad u'(0) = u'(T),$$

has a solution if  $a \in \mathbb{R}$  and  $-\frac{\pi}{2} < \bar{h} < \frac{\pi}{2}$ . This is another example of nonlinear generalization of the resonance condition at zero eigenvalue.

Using topological degree, variational methods, or symplectic techniques, existence results have also been obtained in the case of resonance at a nonzero eigenvalue. For example, the problem

$$u'' + k^2 u + g(u) = h(t), \\ u(0) = u(2\pi), \quad u'(0) = u'(2\pi)$$

with  $g: \mathbb{R} \rightarrow \mathbb{R}$  continuous and bounded and  $k$  a positive integer, has at least one solution for all continuous  $h: [0, 2\pi] \rightarrow \mathbb{R}$  such that the real function

$$\Phi(\theta) := 2 [\widehat{g}(+\infty) - \widehat{g}(-\infty)] - \int_0^{2\pi} h(t) \sin k(t + \theta) dt$$

has no zero or more than two zeros, all simple, in  $[0, 2\pi/k]$ . See [17] for the proof and references to earlier contributions of Lazer–Leach, Dancer, and Fabry–Fonda.

### Future Directions

In the whole XXth century, the study of periodic solutions of non-autonomous ordinary differential equations has been highly influential in the creation and development of fundamental parts of present mathematics, like functional analysis (operator theory, iterative methods, . . .), algebraic topology (fixed point theorems, topological degree, Conley index, . . .), variational methods (dual action principle, minimax theorems, Morse theory, . . .), symplectic techniques (Poincaré–Birkhoff-type fixed point theorems, . . .). One can expect that further topological tools will be useful or developed in searching new existence and multiplicity theorems for periodic solutions.

The difficult problem of discussing the stability of those periodic solutions, still in infancy, is fundamental for the applications and must be developed. Most earlier studies of special differential equations have been devoted to models coming from mechanics and electronics, which are far from being completely understood, but the recent applications to biology, demography and economy will introduce new classes of differential equations and systems

with periodic time dependence, requiring the use of new analytical and topological tools. In this respect, the study of periodic solutions of nonlinear nonautonomous difference equations is of increasing importance.

Even if periodic solutions are not the easiest objects to be found numerically with a large degree of certitude, numerical methods and computers will play an increasing role in detecting the presence of periodic solutions. Some efficient softwares have already been developed in this respect.

An enormous gap still exists between the methods of approach and the results about what seems to be the natural generalization of periodic solutions, namely the almost periodic solutions. It is a paradoxical fact that some statements are true both for periodic solutions and for solutions bounded on the whole real line, but false for the intermediate case of almost periodic solutions!

Finally, like the equilibria in autonomous systems, the study of periodic solutions of non-autonomous equations will remain the unavoidable first step in trying to understand the complexity of the set of all solutions, and much remains to be done in this direction. Poincaré's famous sentence

'What renders these periodic solutions so precious is that they are, so to speak, the only breach through which we may try to penetrate a stronghold previously reputed to be impenetrable'

keeps its full significance in the beginning of the XXIth century.

## Bibliography

### Primary Literature

- Ahmad S, Lazer AC, Paul JL (1976) Elementary critical point theory and perturbations of elliptic boundary value problems. *Indiana Univ Math J* 25:933–944
- Alonso JM, Ortega R (1996) Unbounded solutions of semilinear equations at resonance. *Nonlinearity* 9:1099–1111
- Amann H, Ambrosetti A, Mancini G (1978) Elliptic equations with noninvertible Fredholm linear part and bounded nonlinearities. *Math Z* 158:179–194
- Ambrosetti A, Prodi G (1972) On the inversion of some differentiable mappings with singularities between Banach spaces. *Ann Mat Pura Appl* 93(4):231–246
- Bartsch T, Mawhin J (1991) The Leray–Schauder degree of  $S^1$ -equivariant operators associated to autonomous neutral equations in spaces of periodic functions. *J Differ Equations* 92:90–99
- Campos J, Ortega R (1996) Nonexistence of periodic solutions of a complex Riccati equation. *Differ Integral Equations* 9:247–249
- Capietto A, Mawhin J, Zanolin F (1992) Continuation theorems for periodic perturbations of autonomous systems. *Trans Amer Math Soc* 329:41–72
- Capietto A, Mawhin J, Zanolin F (1995) A continuation theorem for periodic boundary value problems with oscillatory nonlinearities. *Nonlinear Differ Equations Appl* 2:133–163
- Chang KC (1989) On the periodic nonlinearity and the multiplicity of solutions. *Nonlinear Anal* 13:527–537
- Chang KC (1993) *Infinite Dimensional Morse Theory and Multiple Solution Problems*. Birkhäuser, Basel
- Conley C, Zehnder E (1984) Morse type index for flows and periodic solutions for Hamiltonian operator. *Comm Pure Appl Math* 37:207–253
- Dancer EN (1982) On the use of asymptotics in nonlinear boundary value problems. *Ann Mat Pura Appl* 131(4):67–187
- Dancer EN, Ortega R (1994) The index of Lyapunov stable fixed points in two dimensions. *J Dyn Differ Equations* 6:631–637
- Dolph CL (1949) Nonlinear integral equations of Hammerstein type. *Trans Amer Math Soc* 66:289–307
- Duffing G (1918) *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz*. Vieweg, Braunschweig
- Fabry C, Mawhin J, Nkashama M (1986) A multiplicity result for periodic solutions of forced nonlinear second order differential equations. *Bull London Math Soc* 18:173–180
- Fabry C, Mawhin J (2000) Oscillations of a forced asymmetric oscillator at resonance. *Nonlinearity* 13:493–505
- Fučík S (1976) Boundary value problems with jumping nonlinearities. *Časopis Pest Mat* 101:69–87
- Hamel G (1922) Ueber erzwungene Schwingungen bei endlichen Amplituden. *Math Ann* 86:1–13
- Kazdan JL, Warner FW (1975) Remarks on some quasilinear elliptic equations. *Comm Pure Appl Math* 28:567–597
- Knobloch HW (1963) Eine neue Methode zur Approximation periodischer Lösungen von Differentialgleichungen zweiter Ordnung. *Math Z* 82:177–197
- Krasnosel'skii MA (1968) The operator of translation along the trajectories of differential equations. *Amer Math Soc, Providence*
- Lazer AC, Leach DE (1969) Bounded perturbations for forced harmonic oscillators at resonance. *Ann Mat Pura Appl* 82(4):49–68
- Lefschetz S (1943) Existence of periodic solutions of certain differential equations. *Proc Nat Acad Sci USA* 29:29–32
- Levinson N (1943) On the existence of periodic solutions for second order differential equations with a forcing term. *J Math Phys* 22:41–48
- Lichtenstein L (1915) Ueber einige Existenzprobleme der Variationsrechnung. *J Reine Angew Math* 145:24–85
- Ljusternik L, Schnirelmann L (1934) *Méthodes topologiques dans les problèmes variationnels*. Hermann, Paris
- Lloyd NG (1975) On a class of differential equations of Riccati type. *J London Math Soc* 10(2):1–10
- Massera JL (1950) The existence of periodic solutions of systems of differential equations. *Duke Math J* 17:457–475
- Mawhin J (1969) Équations intégrales et solutions périodiques de systèmes différentiels non linéaires. *Bull CI Sci Acad Roy Belgique* 55(5):934–947
- Mawhin J (1976) Contractive mappings and periodically perturbed conservative systems. *Arch Math (Brno)* 12:67–73
- Mawhin J (1979) *Topological Degree and Nonlinear Boundary Value Problems*. CBMS Conf Math No 40. Amer Math Soc, Providence

33. Mawhin J (1983) Points fixes, points critiques et problèmes aux limites. Sémin Math Supérieures No 92, Presses Univ de Montréal, Montréal
34. Mawhin J (1989) Forced second order conservative systems with periodic nonlinearity. *Ann Inst Henri Poincaré Anal non linéaire* 6:415–434
35. Mawhin J (1994) Periodic solutions of some planar non-autonomous polynomial differential equations. *Differ Integral Equations* 7:1055–1061
36. Mawhin J (2004) Global results for the forced pendulum equation. In: Cañada A, Drabek P, Fonda A (eds) *Handbook of Differential Equations. Ordinary Differential Equations*, vol 1. Elsevier, Amsterdam, pp 533–590
37. Mawhin J, Ward JR (2002) Guiding-like functions for periodic or bounded solutions of ordinary differential equations. *Discret Continuous Dyn Syst* 8:39–54
38. Mawhin J, Willem M (1984) Multiple solutions of the periodic boundary value problem for some forced pendulum-type equations. *J Differ Equations* 52:264–287
39. Mawhin J, Willem M (1989) Critical point theory and Hamiltonian systems. Springer, New York
40. Ortega R (1995) Some applications of the topological degree to stability theory. In: Granas A, Frigon M (eds) *Topological methods in differential equations and inclusions*, NATO ASI C472. Kluwer, Amsterdam, pp 377–409
41. Palais R (1966) Ljusternik–Schnirelmann theory on Banach manifolds. *Topology* 5:115–132
42. Palais R, Smale S (1964) A generalized Morse theory. *Bull Amer Math Soc* 70:165–171
43. Poincaré H (1892) Les méthodes nouvelles de la mécanique céleste, vol 1. Gauthier–Villars, Paris
44. Rabinowitz P (1978) Some minimax theorems and applications to nonlinear partial differential equations. In: Cesari L, Kannan R, Weinberger H (eds) *Nonlinear Analysis, A tribute to E. Rothe*. Academic Press, New York
45. Rabinowitz P (1988) On a class of functionals invariant under a  $\mathbb{Z}^n$ -action. *Trans Amer Math Soc* 310:303–311
46. Reissig R (1964) Ein funktionenanalytischer Existenzbeweis für periodische Lösungen. *Monatsber Deutsche Akad Wiss Berlin* 6:407–413
47. Scorza Dragoni G (1931) Il problema dei valori al limite studiate in grande per le equazione differenziale del secondo ordine. *Math Ann* 105:133–143
48. Szrednicki R (1994) On periodic planar solutions of planar polynomial differential equations with periodic coefficients. *J Differ Equations* 114:77–100
49. Stoppelli F (1952) Su un'equazione differenziale della meccanica dei fili. *Rend Accad Sci Fis Mat Napoli* 19(4):109–114
50. Villari G (1965) Contributi allo studio dell'esistenzadi soluzioni periodiche per i sistemi di equazioni differenziali ordinarie. *Ann Mat Pura Appl* 69(4):171–190
51. Willem M (1981) Oscillations forcées de systèmes hamiltoniens. Publications du Séminaire d'Analyse non linéaire, Univ Besançon
- Cesari L (1971) Asymptotic behavior and stability problems for ordinary differential equations, 3rd edn. Springer, Berlin
- Coddington EA, Levinson N (1955) Ordinary differential equations. McGraw-Hill, New York
- Cronin J (1964) Fixed points and topological degree in nonlinear analysis. Amer Math Soc, Providence
- Ekeland I (1990) Convexity methods in hamiltonian mechanics. Springer, Berlin
- Farkas M (1994) Periodic motions. Springer, New York
- Gaines RE, Mawhin J (1977) Coincidence degree and nonlinear differential equations. *Lecture Notes in Mathematics*, vol 568. Springer, Berlin
- Hale JK (1969) Ordinary differential equations. Wiley Interscience, New York
- Mawhin J (1993) Topological degree and boundary value problems for nonlinear differential equations. *Lecture Notes in Mathematics*, vol 1537. Springer, Berlin, pp 74–142
- Mawhin J (1995) Continuation theorems and periodic solutions of ordinary differential equations. In: Granas A, Frigon M (eds) *Topological methods in differential equations and inclusions*, NATO ASI C472. Kluwer, Amsterdam, pp 291–375
- Pliss VA (1966) Nonlocal problems of the theory of oscillations. Academic Press, New York
- Rabinowitz P (1986) Minimax methods in critical point theory with applications to differential equations. *CBMS Reg Conf Ser Math* No 65. Amer Math Soc, Providence
- Reissig R, Sansone G, Conti R (1963) Qualitative Theorie nichtlinearer Differentialgleichungen. Cremonese, Roma
- Reissig R, Sansone G, Conti R (1974) Nonlinear differential equations of higher order. Noordhoff, Leiden
- Rouche N, Mawhin J (1980) Ordinary differential equations. Stability and periodic solutions. Pitman, Boston
- Sansone G, Conti R (1964) Non-linear differential equations. Pergamon, Oxford
- Verhulst F (1996) Nonlinear differential equations and dynamical systems, 2nd edn. Springer, Berlin

## Perturbation Analysis of Parametric Resonance

FERDINAND VERHULST  
Mathematisch Instituut, University of Utrecht,  
Utrecht, The Netherlands

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Perturbation Techniques  
Parametric Excitation of Linear Systems  
Nonlinear Parametric Excitation  
Applications  
Future Directions  
Acknowledgment  
Bibliography

### Books and Reviews

- Bobylev NA, Burman YM, Korovin SK (1994) Approximation procedures in nonlinear oscillation theory. de Gruyter, Berlin
- Bogoliubov NN, Mitropolsky YA (1961) Asymptotic methods in the theory of nonlinear oscillations. Gordon and Breach, New York

## Glossary

**Coexistence** The special case when all the independent solutions of a linear,  $T$ -periodic ODE are  $T$ -periodic.

**Hill's equation** A second order ODE of the form  $\ddot{x} + p(t)x = 0$ , with  $p(t)$   $T$ -periodic.

**Instability pockets** Finite domains, usually intersections of instability tongues, where the trivial solution of linear,  $T$ -periodic ODEs is unstable.

**Instability tongues** Domains in parameter space where the trivial solution of linear,  $T$ -periodic ODEs is unstable.

**Mathieu equation** An ODE of the form  $\ddot{x} + (a + b \cos(t))x = 0$ .

**Parametric resonance** Resonance excitation arising for special values of coefficients, frequencies and other parameters in  $T$ -periodic ODEs.

**Quasi-periodic** A function of the form  $\sum_{i=1}^n f_i(t)$  with  $f_i(t)$   $T_i$ -periodic,  $n$  finite, and the periods  $T_i$  independent over  $\mathbb{R}$ .

**Sum resonance** A parametric resonance arising in the case of at least three frequencies in a  $T$ -periodic ODE.

## Definition of the Subject

Parametric resonance arises in mechanics in systems with external sources of energy and for certain parameter values. Typical examples are the pendulum with oscillating support and a more specific linearization of this pendulum, the Mathieu equation in the form

$$\ddot{x} + (a + b \cos(t))x = 0.$$

The time-dependent term represents the excitation. Tradition has it that parametric resonance is usually not considered in the context of systems with *external* excitation of the form  $\dot{x} = f(x) + \phi(t)$ , but for systems where time-dependence arises in the coefficients of the equation. Mechanically this means usually periodically varying stiffness, mass or load, in fluid or plasma mechanics one can think of frequency modulation or density fluctuation, in mathematical biology of periodic environmental changes. The term 'parametric' refers to the dependence on parameters and certain resonances arising for special values of the parameters. In the case of the Mathieu equation, the parameters are the frequency  $\omega$  ( $a = \omega^2$ ) of the equation without time-dependence and the excitation amplitude  $b$ ; see Sect. "Parametric Excitation of Linear Systems" for an explicit demonstration of resonance phenomena in this two parameters system.

Mathematically the subject is concerned with ODEs with periodic coefficients. The study of linear dynamics of

this type gave rise to a large amount of literature in the first half of the 20th century and this highly technical, classical material is still accessible in textbooks. The standard equations are Hill's equation and the Mathieu equation (see Subsect. "Elementary Theory"). We will summarize a number of basic aspects. The reader is also referred to the article ► [Dynamics of Parametric Excitation](#) by Alan Champneys in this Encyclopedia.

Recently, the interest in nonlinear dynamics, new applications and the need to explore higher dimensional problems has revived the subject. Also structural stability and persistence problems have been investigated. Such problems arise as follows. Suppose that we have found a number of interesting phenomena for a certain equation and suppose we embed this equation in a family of equations by adding parameters or perturbations. Do the 'interesting phenomena' persist in the family of equations? If not, we will call the original equation structurally unstable. A simple example of structural instability is the harmonic equation which shows qualitative different behavior on adding damping. In general, Hamiltonian systems are structurally unstable in the wider context of dissipative dynamical systems.

## Introduction

Parametric resonance produces interesting mathematical challenges and plays an important part in many applications. The linear dynamics is already nontrivial whereas the nonlinear dynamics of such systems is extremely rich and largely unexplored. The role of symmetries is essential, both in linear and in nonlinear analysis. A classical example of parametric excitation is the swinging pendulum with oscillating support. The equation of motion describing the model is

$$\ddot{x} + (\omega_0^2 + p(t)) \sin x = 0, \quad (1)$$

where  $p(t)$  is a periodic function. Upon linearization – replacing  $\sin x$  by  $x$  – we obtain Hill's equation (Subsect. "Elementary Theory"):

$$\ddot{x} + (\omega_0^2 + p(t)) x = 0.$$

This equation was formulated around 1900 in the perturbation theory of periodic solutions in celestial mechanics. If we choose  $p(t) = \cos \omega t$ , Hill's equation becomes the Mathieu equation. It is well-known that special tuning of the frequency  $\omega_0$  and the period of excitation (of  $p(t)$ ) produces interesting instability phenomena (resonance). More generally we may study nonlinear parametric equations of the form

$$\ddot{x} + k\dot{x} + (\omega_0^2 + p(t))F(x) = 0, \quad (2)$$



where  $k > 0$  is the damping coefficient,  $F(x) = x + bx^2 + cx^3 + \dots$  and time is scaled so that  $p(t)$  is a  $\pi$ -periodic function with zero average. We may also take for  $p(t)$  a quasi-periodic or almost-periodic function. The books [48] cover most of the classical theory, but for a nice introduction see [38]. In [36], emphasis is placed on the part played by parameters, it contains a rich survey of bifurcations of eigenvalues and various applications. There are many open questions for Eqs. (1) and (2); we shall discuss aspects of the classical theory, recent theoretical results and a few applications.

As noted before, in parametric excitation we have an oscillator with an independent source of energy. In examples, the oscillator is often described by a one degree of freedom system but of course many more degrees of freedom may play a part; see for instance in Sect. “Applications” the case of coupled Mathieu-equations as studied in [33]. In what follows,  $\varepsilon$  will always be a small, positive parameter.

## Perturbation Techniques

In this section we review the basic techniques to handle parametric perturbation problems. In the case of Poincaré–Lindstedt series which apply to periodic solutions, the expansions are in integer powers of  $\varepsilon$ . It should be noted that in general, other order functions of  $\varepsilon$  may play a part; see Subsect. “Elementary Theory” and [46].

### Poincaré–Lindstedt Series

One of the oldest techniques is to approximate a periodic solution by the construction of a convergent series in terms of the small parameter  $\varepsilon$ . The method can be used for equations of the form

$$\dot{x} = f(t, x) + \varepsilon g(t, x) + \varepsilon^2 \dots,$$

with  $x \in \mathbb{R}^n$  and (usually) assuming that the ‘unperturbed’ problem  $\dot{y} = f(t, y)$  is understood and can be solved. Note that the method can also be applied to perturbed maps and difference equations. Suppose that the unperturbed problem contains a periodic solution, under what conditions can this solution be continued for  $\varepsilon > 0$ ? The answer is given by the conditions set by the implicit function theorem, see for formulations and theorems [30] and [44]. Usually we can associate with our perturbation problem a parameter space and one of the questions is then to find the domains of stability and instability. The common boundary of these domains is often characterized by the existence of periodic solutions and this is where Poincaré–Lindstedt series are useful. We will demonstrate this in the next section.

### Averaging

Averaging is a normalization method. In general, the term “normalization” is used whenever an expression or quantity is put in a simpler, standardized form. For instance, a  $n \times n$ -matrix with constant coefficients can be put in Jordan normal form by a suitable transformation. When the eigenvalues are distinct, this is a diagonal matrix.

Introductions to normalization can be found in [1, 15, 19] and [13]. For the relation between averaging and normalization in general the reader is referred to [34] and [44]. For averaging in the so-called standard form it is assumed that we can put the perturbation problem in the form

$$\dot{x} = \varepsilon F(t, x) + \varepsilon^2 \dots,$$

and that we have the existence of the limit

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T F(t, x) dt = F^0(x).$$

The analysis of the averaged equation  $\dot{y} = F^0(y)$  produces asymptotic approximations of the solutions of the original equation on a long timescale; see [34]. Also, under certain conditions, critical points of the averaged equation correspond with periodic solutions in the original system. The choice to use Poincaré–Lindstedt series or the averaging method is determined by the amount of information one wishes to obtain. To find the location of stability and instability domains (the boundaries), Poincaré–Lindstedt series are very efficient. On the other hand, with somewhat more efforts, the averaging method will also supply this information with in addition the behavior of the solutions within the domains. For an illustration see Subsect. “Elementary Theory”.

### Resonance

Assume that  $x = 0$  is a critical point of the differential equation and write the system as:

$$\dot{x} = Ax + f(t, x, \varepsilon), \quad (3)$$

with  $x \in \mathbb{R}^n$ ,  $A$  a constant  $n \times n$ -matrix;  $f(t, x, \varepsilon)$  can be expanded in a Taylor series with respect to  $\varepsilon$  and in homogeneous vector polynomials in  $x$  starting with quadratic terms. Normalization of Eq. (3) means that by successive transformation we remove as many terms of Eq. (3) as possible. It would be ideal if we could remove all the nonlinear terms, i. e. linearize Eq. (3) by transformation. In general, however, some nonlinearities will be left and this is where

resonance comes in. The eigenvalues  $\lambda_1, \dots, \lambda_n$  of the matrix  $A$  are resonant if for some  $i \in \{1, 2, \dots, n\}$  one has:

$$\sum_{j=1}^n m_j \lambda_j = \lambda_i, \quad (4)$$

with  $m_j \geq 0$  integers and  $m_1 + m_2 + \dots + m_n \geq 2$ . If the eigenvalues of  $A$  are non-resonant, we can remove all the nonlinear terms and so linearize the system. However, this is less useful than it appears, as in general the sequence of successive transformations to carry out the normalization will be divergent. The usefulness of normalization lies in removing nonresonant terms to a certain degree to simplify the analysis.

### Normalization of Time-Dependent Vectorfields

In problems involving parametric resonance, we have time-dependent systems such as equations perturbing the Mathieu equation. Details of proofs and methods to compute the normal form coefficients in such cases can be found in [1,18] and [34]. We summarize some aspects. Consider the following parameter and time dependent equation:

$$\dot{x} = F(x, \mu, t), \quad (5)$$

with  $x \in \mathbb{R}^m$  and the parameters  $\mu \in \mathbb{R}^p$ .

Here  $F(x, \mu, t) : \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \rightarrow \mathbb{R}^m$  is  $C^\infty$  in  $x$  and  $\mu$  and  $T$ -periodic in the  $t$ -variable. We assume that  $x = 0$  is a solution, so  $F(0, \mu, t) = 0$  and, moreover assume that the linear part of the vectorfield  $D_x F(0, 0, t)$  is time-independent for all  $t \in \mathbb{R}$ . We will write  $L_0 = D_x F(0, 0, t)$ . Expanding  $F(x, \mu, t)$  in a Taylor series with respect to  $x$  and  $\mu$  yields the equation:

$$\dot{x} = L_0 x + \sum_{n=2}^k F_n(x, \mu, t) + O(|(x, \mu)|^{k+1}), \quad (6)$$

where the  $F_n(x, \mu, t)$  are homogeneous polynomials in  $x$  and  $\mu$  of degree  $n$  with  $T$ -periodic coefficients.

**Theorem 1** *Let  $k \in \mathbb{N}$ . There exists a (parameter- and time-dependent) transformation*

$$x = \hat{x} + \sum_{n=2}^k P_n(\hat{x}, \mu, t),$$

where  $P_n(\hat{x}, \mu, t)$  are homogeneous polynomials in  $x$  and  $\mu$  of degree  $n$  with  $T$ -periodic coefficients, such that Eq. (6) takes the form (dropping the hat):

$$\begin{aligned} \dot{x} &= L_0 x + \sum_{n=2}^k \tilde{F}_n(x, \mu, t) + O(|(x, \mu)|^{k+1}), \\ \dot{\mu} &= 0. \end{aligned} \quad (7)$$

The truncated vectorfield:

$$\dot{x} = L_0 x + \sum_{n=2}^k \tilde{F}_n(x, \mu, t) = \tilde{F}(x, \mu, t), \quad (8)$$

which will be called the normal form of Eq. (5), has the following properties:

1.  $\frac{d}{dt} e^{L_0^* t} \tilde{F}(e^{-L_0^* t} x, \mu, t) = 0$ , for all  $(x, \mu) \in \mathbb{R}^{m+p}$ ,  $t \in \mathbb{R}$ .
2. If Eq. (5) is invariant under an involution (i.e.  $SF(x, \mu, t) = F(Sx, \mu, t)$  with  $S$  an invertible linear operator such that  $S^2 = I$ ), then the truncated normal form (8) is also invariant under  $S$ . Similarly, if Eq. (5) is reversible under an involution  $R$  (i.e.  $RF(x, \mu, t) = -F(Rx, \mu, t)$ ), then the truncated normal form (8) is also reversible under  $R$ .

For a proof, see [18].

The theorem will be applied to situations where  $L_0$  is semi-simple and has only purely imaginary eigenvalues. We take  $L_0 = \text{diag}\{i\lambda_1, \dots, i\lambda_m\}$ . In our applications,  $m = 2l$  is even and  $\lambda_{l+j} = -\lambda_j$  for  $j = 1, \dots, l$ . The variable  $x$  is then often written as  $x = (z_1, \dots, z_l, \bar{z}_1, \dots, \bar{z}_l)$ .

Assume  $L_0 = \text{diag}\{i\lambda_1, \dots, i\lambda_m\}$  then:

- A term  $x_1^{\gamma_1} \dots x_m^{\gamma_m} e^{i \frac{2\pi}{T} k t}$  is in the  $j$ th component of the Taylor-Fourier series of  $\tilde{F}(x, \mu, t)$  if:

$$-\lambda_j + \frac{2\pi}{T} k + \gamma_1 \lambda_1 + \dots + \gamma_m \lambda_m = 0. \quad (9)$$

This is known as the resonance condition.

- Transforming the normal form through  $x = e^{L_0 t} w$  leads to an autonomous equation for  $w$ :

$$\dot{w} = \sum_{n=2}^k \tilde{F}_n(w, \mu, 0). \quad (10)$$

- An important result is this: If Eq. (5) is invariant (respectively reversible) under an involution  $S$ , then this also holds for Eq. (10).
- The autonomous normal form (10) is invariant under the action of the group  $\bar{G} = \{g|gx = e^{jL_0 T} x, j \in \mathbb{Z}\}$ , generated by  $e^{L_0 T}$ . Note that this group is discrete if the ratios of the  $\lambda_i$  are rational and continuous otherwise.

For a proof of the last two statements see [31].

By this procedure we can make the system autonomous. This is very effective as the autonomous normal form (10) can be used to prove the existence of periodic solutions and invariant tori of Eq. (5) near  $x = 0$ . We have:

**Theorem 2** Let  $\varepsilon > 0$ , sufficiently small, be given. Scale  $w = \varepsilon \dot{w}$ .

1. If  $\dot{w}_0$  is a hyperbolic fixed point of the (scaled) Eq. (10), then Eq. (5) has a hyperbolic periodic solution  $x(t) = \varepsilon \dot{w}_0 + O(\varepsilon^{k+1})$ .
2. If the scaled Eq. (10) has a hyperbolic closed orbit, then Eq. (5) has a hyperbolic invariant torus.

These results are related to earlier theorems in [3], see also the survey [45]. Later we shall discuss normalization in the context of the so-called sum-resonance.

### Remarks on Limit Sets

In studying a dynamical system the behavior of the solutions is for a large part determined by the limit sets of the system. The classical limit sets are equilibria and periodic orbits. Even when restricting to autonomous equations of dimension three, we have no complete classification of possible limit sets and this makes the recognition and description of non-classical limit sets important. In parametrically excited systems, the following limit sets, apart from the classical ones, are of interest:

- Chaotic attractors. Various scenarios were found, see [31,32,42].
- Strange attractors without chaos, see [27]. The natural presence of various forcing periods in real-life models make their occurrence quite plausible.
- Attracting tori. These limit sets are not difficult to find; they arise for instance as a consequence of a Neimark–Sacker bifurcation of a periodic solution, see [19].
- Attracting heteroclinic cycles, see [20].

A large number of these phenomena can be studied both by numerics and by perturbation theory; using the methods simultaneously gives additional insight.

### Parametric Excitation of Linear Systems

As we have seen in the introduction, parametric excitation leads to the study of second order equations with periodic coefficients. More in general such equations arise from linearization near  $T$ -periodic solutions of  $T$ -periodic equations of the form  $\dot{y} = f(t, y)$ . Suppose  $y = \phi(t)$  is a  $T$ -periodic solution; putting  $y = \phi(t) + x$  produces upon linearization the  $T$ -periodic equation

$$\dot{x} = f_x(t, \phi(t))x. \quad (11)$$

This equation often takes the form

$$\dot{x} = Ax + \varepsilon B(t)x, \quad (12)$$

in which  $x \in \mathbb{R}^m$ ;  $A$  is a constant  $m \times m$ -matrix,  $B(t)$  is a continuous,  $T$ -periodic  $m \times m$ -matrix,  $\varepsilon$  is a small parameter. For elementary studies of such an equation, the Poincaré–Lindstedt method or continuation method is quite efficient. The method applies to nonlinear equations of arbitrary dimension, but we shall demonstrate its use for equations of Mathieu type.

### Elementary Theory

Floquet theory tells us that the solutions of Eq. (12) can be written as:

$$x(t) = \Phi(t, \varepsilon)e^{C(\varepsilon)t}, \quad (13)$$

with  $\Phi(t, \varepsilon)$  a  $T$ -periodic  $m \times m$ -matrix,  $C(\varepsilon)$  a constant  $m \times m$ -matrix and both matrices having an expansion in order functions of  $\varepsilon$ . The determination of  $C(\varepsilon)$  provides us with the stability behavior of the solutions. A particular case of Eq. (12) is Hill's equation:

$$\ddot{x} + b(t, \varepsilon)x = 0, \quad (14)$$

which is of second order;  $b(t, \varepsilon)$  is a scalar  $T$ -periodic function. A number of cases of Hill's equation are studied in [23]. A particular case of Eq. (14) which arises frequently in applications is the Mathieu equation:

$$\ddot{x} + (\omega^2 + \varepsilon \cos 2t)x = 0, \omega > 0, \quad (15)$$

which is reversible. (In [23] one also finds Lamé's, Ince's, Hermite's, Whittaker–Hill and other Hill equations.) A typical question is: for which values of  $\omega$  and  $\varepsilon$  in  $(\omega^2, \varepsilon)$ -parameter space is the trivial solution  $x = \dot{x} = 0$  stable?

Solutions of Eq. (15) can be written in the Floquet form (13), where in this case  $\Phi(t, \varepsilon)$  will be  $\pi$ -periodic. The eigenvalues  $\lambda_1, \lambda_2$  of  $C$ , which are called characteristic exponents and are  $\varepsilon$ -dependent, determine the stability of the trivial solution. For the characteristic exponents of Eq. (12) we have:

$$\sum_{i=1}^n \lambda_i = \frac{1}{T} \int_0^T \text{Tr}(A + \varepsilon B(t))dt, \quad (16)$$

see Theorem 6.6 in [44]. So in the case of Eq. (15) we have:

$$\lambda_1 + \lambda_2 = 0. \quad (17)$$

The exponents are functions of  $\varepsilon$ ,  $\lambda_1 = \lambda_1(\varepsilon)$ ,  $\lambda_2 = \lambda_2(\varepsilon)$  and clearly  $\lambda_1(0) = i\omega$ ,  $\lambda_2(0) = -i\omega$ . As  $\lambda_1(\varepsilon) = -\lambda_2(\varepsilon)$ , the characteristic exponents, which are complex conjugate, are purely imaginary or real. The implication is

that if  $\omega^2 \neq n^2$ ,  $n = 1, 2, \dots$  the characteristic exponents are purely imaginary and  $x = 0$  is stable near  $\varepsilon = 0$ . If  $\omega^2 = n^2$  for some  $n \in \mathbb{N}$ , however, the imaginary part of  $\exp(C(\varepsilon)t)$  can be absorbed into  $\Phi(t, \varepsilon)$  and the characteristic exponents may be real. We assume now that  $\omega^2 = n^2$  for some  $n \in \mathbb{N}$ , or near this value, and we shall look for periodic solutions of  $x(t)$  of Eq. (15) as these solutions define the boundaries between stable and unstable solutions. We put:

$$\omega^2 = n^2 - \varepsilon\beta, \quad (18)$$

with  $\beta$  a constant, and we apply the Poincaré-Lindstedt method to find the periodic solutions; see Appendix 2 in [44]. We find that periodic solutions exist for  $n = 1$  if:

$$\omega^2 = 1 \pm \frac{1}{2}\varepsilon + \mathcal{O}(\varepsilon^2).$$

In the case  $n = 2$ , periodic solutions exist if:

$$\begin{aligned} \omega^2 &= 4 - \frac{1}{48}\varepsilon^2 + \mathcal{O}(\varepsilon^4), \\ \omega^2 &= 4 + \frac{5}{48}\varepsilon^2 + \mathcal{O}(\varepsilon^4). \end{aligned} \quad (19)$$

The corresponding instability domains are called Floquet tongues, instability tongues or resonance tongues, see Fig. 1.

On considering higher values of  $n$ , we have to calculate to a higher order of  $\varepsilon$ . At  $n = 1$  the boundary curves are intersecting at positive angles at  $\varepsilon = 0$ , at  $n = 2$  ( $\omega^2 = 4$ )



**Perturbation Analysis of Parametric Resonance, Figure 1**  
Floquet tongues of the Mathieu Eq. (15); the instability domains are shaded

they are tangent; the order of tangency increases as  $n - 1$  (contact of order  $n$ ), making instability domains more and more narrow with increasing resonance number  $n$ .

**Higher Order Approximation and an Unexpected Timescale** The instability tongue of the Mathieu equation at  $n = 1$  can be determined with more precision by Poincaré expansion. On using averaging, one also characterizes the flow outside the tongue boundary and this results in a surprise. Consider Eq. (15) in the form

$$\ddot{x} + (1 + \varepsilon a + \varepsilon^2 b + \varepsilon \cos 2t)x = 0,$$

where we can choose  $a = \pm \frac{1}{2}$  to put the frequency with first order precision at the tongue boundary. The eigenvalues of the trivial solution are from first order averaging

$$\lambda_{1,2} = \pm \frac{1}{2} \sqrt{\frac{1}{4} - a^2},$$

which agrees with Poincaré expansion;  $a^2 > \frac{1}{4}$  gives stability, the  $<$  inequality instability. The transition value  $a^2 = \frac{1}{4}$  gives the tongue location. Take for instance the  $+$  sign. Second order averaging, see [46], produces for the eigenvalues of the trivial solution

$$\lambda_{1,2} = \pm \sqrt{-\frac{1}{4} \left( \frac{1}{32} + b \right) \varepsilon^3 + \left( \frac{1}{64} + \frac{1}{2}b \right) \cdot \left( \frac{7}{64} - \frac{1}{2}b \right) \varepsilon^4}.$$

So, if  $\frac{1}{32} + b > 0$  we have stability, if  $\frac{1}{32} + b < 0$  instability; at  $b = -\frac{1}{32}$  we have the second order approximation of this tongue boundary. Note that near this boundary the solutions are characterized by eigenvalues of  $\mathcal{O}(\varepsilon^{\frac{3}{2}})$  and accordingly the time-dependence by timescale  $\varepsilon^{\frac{3}{2}}t$ .

**The Mathieu Equation with Viscous Damping** In real-life applications there is always the presence of damping. We shall consider the effect of its simplest form, small viscous damping. Eq. (15) is extended by adding a linear damping term:

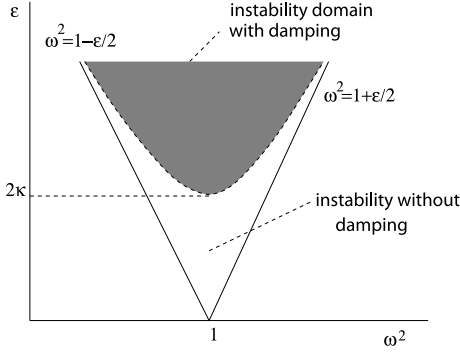
$$\ddot{x} + \kappa \dot{x} + (\omega^2 + \varepsilon \cos 2t)x = 0, \quad a, \kappa > 0. \quad (20)$$

We assume that the damping coefficient is small,  $\kappa = \varepsilon\kappa_0$ , and again we put  $\omega^2 = n^2 - \varepsilon\beta$  to apply the Poincaré-Lindstedt method.

We find periodic solutions in the case  $n = 1$  if:

$$\omega^2 = 1 \pm \sqrt{\frac{1}{4}\varepsilon^2 - \kappa^2}. \quad (21)$$

Relation (21) corresponds with the curve of periodic solutions, which in  $(\omega^2, \varepsilon)$ -parameter space separates stable and unstable solutions. We observe the following phenomena. If  $0 < \kappa < \frac{1}{2}\varepsilon$ , we have an instability domain



**Perturbation Analysis of Parametric Resonance, Figure 2**

First order approximation of instability domains without and with damping for Eq. (20) near  $\omega^2 = 1$

which by damping has been lifted from the  $\omega^2$ -axis; also the width has shrunk. If  $\kappa > \frac{1}{2}\epsilon$  the instability domain has vanished. For an illustration see Fig. 2.

Repeating the calculations for  $n \geq 2$  we find no instability domains at all; damping of  $O(\epsilon)$  stabilizes the system for  $\epsilon$  small. To find an instability domain we have to decrease the damping, for instance if  $n = 2$  we have to take  $\kappa = \epsilon^2 \kappa_0$ .

**Coexistence** Linear periodic equations of the form (12) have  $m$  independent solutions and it is possible that all the independent solutions are periodic. This is called ‘coexistence’ and one of the consequences is that the instability tongues vanish. An example is Ince’s equation:

$$(1 + a \cos t)\ddot{x} + \kappa \sin t \dot{x} + (\omega^2 + \epsilon \cos t)x = 0,$$

see [23]. An interesting question is whether this phenomenon persists under nonlinear perturbations; we return to this question in Subsect. “Coexistence Under Nonlinear Perturbation”.

### More General Classical Results

The picture presented by the Mathieu equation resulting in resonance tongues in the  $\omega, \epsilon$ -parameter space, stability and instability intervals as parametrized by  $\omega$  shown in Fig. 1, has been studied for more general types of Hill’s equation. The older literature can be found in [39], see also [43].

Consider Hill’s equation in the form

$$\ddot{x} + (\omega^2 + \epsilon f(t))x = 0, \quad (22)$$

with  $f(t)$  periodic and represented by a Fourier series. Along the  $\omega^2$ -axis there exist instability intervals of size

$L_m$ , where  $m$  indicates the  $m$ th instability interval. In the case of the Mathieu equation, we have from [16]

$$L_m = O(\epsilon^m).$$

The resonance tongues become increasingly narrow.

For general periodic  $f(t)$  we have weak estimates, like  $L_m = O(\epsilon)$ , but if we assume that the Fourier series is finite, the estimates can be improved. Put

$$f(t) = \sum_{j=0}^s f_j \cos 2jt,$$

so  $f(t)$  is even and  $\pi$ -periodic. From [22] we have the following estimates:

- If we can write  $m = sp$  with  $p \in \mathbb{N}$ , we have

$$L_m = \frac{8s^2}{((p-1)!)^2} \left( \frac{|f_s \epsilon|}{8s^2} \right)^p + O(\epsilon^{p+1}).$$

- If we can not decompose  $m$  like this and  $sp < m < s(p+1)$ , we have

$$L_m = O(\epsilon^{p+1}).$$

In the case of Eq. (22) we have no dissipation and then it can be useful to introduce canonical transformations and Poincaré maps. In this case, for example, put

$$\dot{x} = y, \quad \dot{y} = -\frac{\partial H}{\partial x},$$

with Hamiltonian function

$$H(x, y, t) = \frac{1}{2}y^2 + \frac{1}{2}(\omega^2 + \epsilon f(t))x^2.$$

We can split  $H = H_0 + \epsilon H_1$  with  $H_0 = \frac{1}{2}(y^2 + \omega^2 x^2)$  and apply canonical perturbation theory. Examples of this line of research can be found in [6] and [10]. Interesting conclusions can be drawn with respect to the geometry of the resonance tongues, crossings of tongues and as a possible consequence the presence of so-called instability pockets. In this context, the classical Mathieu equation turns out to be quite degenerate.

Hill’s equation in the case of damping was considered in [35]; see also [36] where an arbitrary number of degrees of freedom is discussed.

### Quasi-Periodic Excitation

Equations of the form

$$\ddot{x} + (\omega^2 + \epsilon p(t))x = 0, \quad (23)$$



with parametric excitation  $p(t)$  quasi-periodic or almost-periodic, arise often in applications. Floquet theory does not apply in this case but we can still use perturbation theory. A typical example would be two rationally independent frequencies:

$$p(t) = \cos t + \cos \gamma t,$$

with  $\gamma$  irrational. As an interesting example, in [7],  $\gamma = \frac{1}{2}(1 + \sqrt{5})$  was chosen, the golden number. It will be no surprise that many more complications arise for large values of  $\varepsilon$ , but for  $\varepsilon$  small (the assumption in this article), the analysis runs along similar lines producing resonance tongues, crossings of tongues and instability pockets. See also extensions in [11]. Detailed perturbation expansions are presented in [49] with a comparison of Poincaré expansion, the harmonic balance method and numerics; there is good agreement between the methods. Real-life models contain dissipation which inspired the authors of [49] to consider the equation

$$\ddot{x} + 2\mu\dot{x} + (\omega^2 + \varepsilon(\cos t + \cos \gamma t))x = 0, \quad \mu > 0,$$

$\gamma$  irrational. They conclude that

- The instability tongues become thinner and recede into the  $\omega$ -axis as  $\mu$  increases.
- High-order resonance tongues seem to be more affected by dissipation than low-order ones producing a dramatic loss of ‘fine detail’, even for small  $\mu$ .
- The results of varying the parameter  $\mu$  certainly needs more investigation.

### Parametrically Forced Oscillators in Sum Resonance

In applications where more than one degree of freedom plays a part, many more resonances are possible. For a number of interesting cases and additional literature see [36]. An important case is the so-called sum resonance. In [17] a geometrical explanation is presented for the phenomena in this case using ‘all’ the parameters as unfolding parameters. It will turn out that four parameters are needed to give a complete description. Fortunately three suffice to visualize the situation. Consider the following type of differential equation with three frequencies

$$\dot{z} = Az + \varepsilon f(z, \omega_0 t; \lambda), \quad z \in \mathbb{R}^4, \quad \lambda \in \mathbb{R}^p, \quad (24)$$

which describes a system of two parametrically forced coupled oscillators. Here  $A$  is a  $4 \times 4$  matrix, containing parameters, and with purely imaginary eigenvalues  $\pm i\omega_1$  and  $\pm i\omega_2$ . The vector valued function  $f$  is  $2\pi$ -periodic in  $\omega_0 t$  and  $f(0, \omega_0 t; \lambda) = 0$  for all  $t$  and  $\lambda$ . Equation (24) can

be resonant in many different ways. We consider the sum resonance

$$\omega_1 + \omega_2 = \omega_0,$$

where the system may exhibit instability. The parameter  $\lambda$  is used to control detuning  $\delta = (\delta_1, \delta_2)$  of the frequencies  $(\omega_1, \omega_2)$  from resonance and damping  $\mu = (\mu_1, \mu_2)$ . We summarize the analysis from [17].

- The first step is to put Eq. (24) into normal form by normalization or averaging. In the normalized equation the time-dependence appears only in the higher order terms. But the autonomous part of this equation contains enough information to determine the stability regions of the origin. The linear part of the normal form is  $\dot{z} = A(\delta, \mu)z$  with

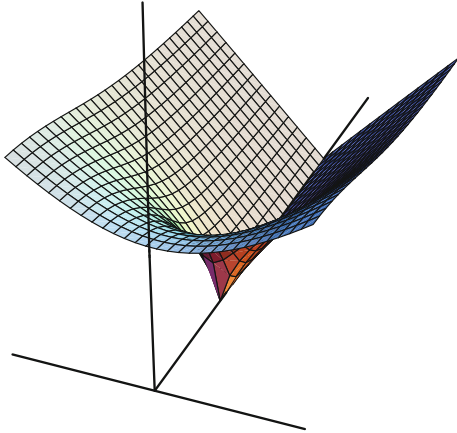
$$A(\delta, \mu) = \begin{pmatrix} B(\delta, \mu) & 0 \\ 0 & \bar{B}(\delta, \mu) \end{pmatrix}, \quad (25)$$

and

$$B(\delta, \mu) = \begin{pmatrix} i\delta_1 - \mu_1 & \alpha_1 \\ \bar{\alpha}_2 & -i\delta_2 - \mu_2 \end{pmatrix}. \quad (26)$$

Since  $A(\delta, \mu)$  is the complexification of a real matrix, it commutes with complex conjugation. Furthermore, according to the normal form Theorem 1 and if  $\omega_1$  and  $\omega_2$  are independent over the integers, the normal form of Eq. (24) has a continuous symmetry group.

- The second step is to test the linear part  $A(\delta, \mu)$  of the normalized equation for structural stability i.e. to answer the question whether there exist open sets in parameter space where the dynamics is qualitatively the same. The family of matrices  $A(\delta, \mu)$  is parametrized by the detuning  $\delta$  and the damping  $\mu$ . We first identify the most degenerate member  $N$  of this family and then show that  $A(\delta, \mu)$  is its versal unfolding in the sense of [1]. The family  $A(\delta, \mu)$  is equivalent to a versal unfolding  $U(\lambda)$  of the degenerate member  $N$ .
- Put differently, the family  $A(\delta, \mu)$  is structurally stable for  $\delta, \mu > 0$ , whereas  $A(\delta, 0)$  is not. This has interesting consequences in applications as small damping and zero damping may exhibit very different behavior, see Sect. “Rotor Dynamics”. In parameter space, the stability regions of the trivial solution are separated by a *critical surface* which is the hypersurface where  $A(\delta, \mu)$  has at least one pair of purely imaginary complex conjugate eigenvalues. This critical surface is diffeomorphic to the *Whitney umbrella*, see Fig. 3 and for references [17]. It is the singularity of the Whitney umbrella that causes the discontinuous behavior of the stability diagram in



**Perturbation Analysis of Parametric Resonance, Figure 3**

The critical surface in  $(\mu_+, \mu_-, \delta_+)$  space for Eq. (24).  $\mu_+ = \mu_1 + \mu_2$ ,  $\mu_- = \mu_1 - \mu_2$ ,  $\delta_+ = \delta_1 + \delta_2$ . Only the part  $\mu_+ > 0$  and  $\delta_+ > 0$  is shown. The parameters  $\delta_1, \delta_2$  control the detuning of the frequencies, the parameters  $\mu_1, \mu_2$  the damping of the oscillators (vertical direction). The base of the umbrella lies along the  $\delta_+$ -axis

Sect. “Rotor Dynamics”. The structural stability argument guarantees that the results are ‘universally valid’, i. e. they qualitatively hold for generic systems in sum resonance.

### Nonlinear Parametric Excitation

Adding nonlinear effects to parametric excitation strongly complicates the dynamics. We start with adding nonlinear terms to the (generalized) Mathieu equation. Consider the following equation that includes dissipation:

$$\ddot{x} + \kappa \dot{x} + (\omega^2 + \varepsilon p(t))f(x) = 0, \quad (27)$$

where  $\kappa > 0$  is the damping coefficient,  $f(x) = x + bx^2 + cx^3 + \dots$ , and time is scaled so that:

$$p(t) = \sum_{l \in \mathbb{Z}} a_{2l} e^{2ilt}, \quad a_0 = 0, \quad a_{-2l} = \bar{a}_{2l}, \quad (28)$$

is an even  $\pi$ -periodic function with zero average. As we have seen in Sect. “Parametric Excitation of Linear Systems”, the trivial solution  $x = 0$  is unstable when  $\kappa = 0$  and  $\omega^2 = n^2$ , for all  $n \in \mathbb{N}$ . Fix a specific  $n \in \mathbb{N}$  and assume that  $\omega^2$  is close to  $n^2$ . We will study the bifurcations from the solution  $x = 0$  in the case of primary resonance, which by definition occurs when the Fourier expansion of  $p(t)$  contains nonzero terms  $a_{2n} e^{2int}$  and  $a_{-2n} e^{-2int}$ . The bifurcation parameters in this problem are the detuning  $\sigma = \omega^2 - n^2$ , the damping coefficient  $\kappa$  and the Fourier coefficients of  $p(t)$ , in particular  $a_{2n}$ . The Fourier coefficients are assumed to be of equal order of magnitude.

### The Conservative Case, $\kappa = 0$

An early paper is [21] in which Eq. (27) for  $\kappa = 0$  is associated with the Hamiltonian

$$H(x, \dot{x}, t) = \frac{1}{2} \dot{x}^2 + \frac{\omega^2}{2} x^2 + p(t) \int_0^x f(s) ds.$$

After transformation of the Hamiltonian, Lie transforms are implemented by MACSYMA to produce normal form approximations to  $O(\varepsilon^2)$ . A number of examples show interesting bifurcations.

A related approach can be found in [5]; as  $p(t)$  is even, the equation is time-reversible. After construction of the Poincaré (timeperiodic) map, normal forms are obtained by equivariant transformations. This leads to a classification of integrable normal forms that are approximations of the family of Poincaré maps, a family as the map is parametrized by  $\omega$  and the coefficients of  $p(t)$ .

Interestingly, the nonlinearity  $\alpha x^3$  is combined with the quasi-periodic Mathieu equation in [50] where global phenomena are described like resonance bands and chaos.

### Adding Dissipation, $\kappa > 0$

Again time-periodic normal form calculations are used to approximate the dynamics; see [31], also [32] and the monograph [42]. The reflection symmetry in the normal form equations implies that all fixed points come in pairs, and that bifurcations of the origin will be symmetric (such as pitchfork bifurcations). We observe that the normal form equations show additional symmetries if either  $f(x)$  in Eq. (27) is odd in  $x$  or if  $n$  is odd. The general normal form can be seen as a non-symmetric perturbation of the symmetric case. One finds pitchfork and saddle-node bifurcations, in fact all codimension one bifurcations; for details and pictures see Chap. 9 in [42].

### Coexistence Under Nonlinear Perturbation

A model describing free vibrations of an elastica is described in [26]:

$$\left(1 - \frac{\varepsilon}{2} \cos 2t\right) \ddot{x} + \varepsilon \sin 2t \dot{x} + cx + \varepsilon \alpha x^2 = 0.$$

For  $\alpha = 0$ , the equation shows the phenomenon of coexistence. It is shown by second order averaging in [26] that for  $\alpha \neq 0$  there exist open sets of parameter values for which the trivial solution is unstable.

An application to the stability problem of a family of periodic solutions in a Hamiltonian system is given in [29].

### Other Nonlinearities

In applications various nonlinear terms play a part. In [25] one considers

$$\ddot{x} + (\omega^2 + \varepsilon \cos(t)) + \varepsilon(Ax^3 + Bx^2\dot{x} + Cx\dot{x}^2 + D\dot{x}^3) = 0,$$

where averaging is applied near the 2 : 1-resonance. If  $B, D < 0$  the corresponding terms can be interpreted as progressive damping. It turns out that for a correct description of the bifurcations second-order averaging is needed.

Nonlinear damping can be of practical interest. The equation

$$\ddot{x} + (\omega^2 + \varepsilon \cos(t)) + \mu|\dot{x}|\dot{x} = 0,$$

is studied with  $\mu$  also a small parameter. A special feature is that an acceptable description of the phenomena can be obtained in a semi-analytical way by using Mathieu-functions as starting point. The analysis involves the use of Padé-approximants, see [28].

### Applications

There are many applications of parametric resonance, in particular in engineering. In this section we consider a number of significant applications, but of course without any attempt at completeness. See also [36] and the references in the additional literature.

#### The Parametrically Excited Pendulum

Choosing the pendulum case  $f(x) = \sin(x)$  in Eq. (27) we have

$$\ddot{x} + \kappa\dot{x} + (\omega^2 + \varepsilon p(t)) \sin(x) = 0.$$

It is natural, because of the sin periodicity, to analyze the Poincaré map on the cylindrical section  $t = 0 \bmod 2\pi\mathbb{Z}$ . This map has both a spatial and a temporal symmetry. As we know from the preceding section, perturbation theory applied near the equilibria  $x = 0, x = \pi$ , produces integrable normal forms. For larger excitation (larger values of  $\varepsilon$ ), the system exhibits the usual picture of Hamiltonian chaos; for details see [12,24].

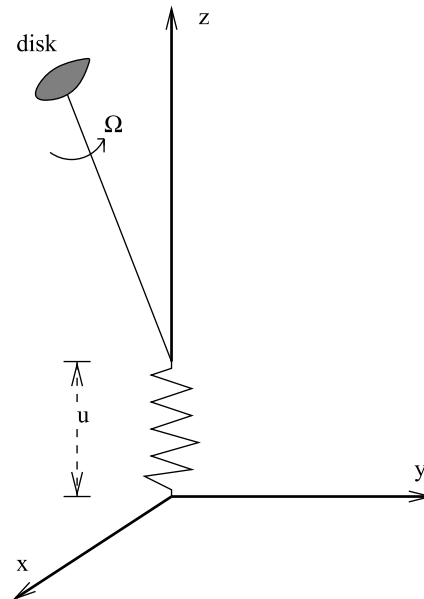
The inverted case is intriguing. It is well-known that the upper equilibrium of an undamped pendulum can be stabilized by vertical oscillations of the suspension point with certain frequencies. See for references [8,9] and [37]. In [8] the genericity of the classical result is studied for (conservative) perturbations respecting the symmetries of

the equation. In [9] genericity is studied for (conservative) perturbations where the spatial symmetry is broken, replacing  $\sin x$  by more general  $2\pi$ -periodic functions. Stabilization is still possible but the dynamics is more complicated.

### Rotor Dynamics

When adding linear damping to a system there can be a striking discontinuity in the bifurcational behavior. Phenomena like this have already been observed and described in for instance [48] or [40]. The discontinuity is a fundamental structural instability in linear gyroscopic systems with at least two degrees of freedom and with linear damping. The following example is based on [41] and [33].

Consider a rigid rotor consisting of a heavy disk of mass  $M$  which is rotating with rotationspeed  $\Omega$  around an axis. The axis of rotation is elastically mounted on a foundation; the connections which are holding the rotor in an upright position are also elastic. To describe the position of the rotor we have the axial displacement  $u$  in the vertical direction (positive upwards), the angle of the axis of rotation *with respect to the z-axis* and *around the z-axis*. Instead of these two angles we will use the projection of the center of gravity motion on the horizontal  $(x, y)$ -plane, see Fig. 4. Assuming small oscillations in the upright ( $u$ ) posi-



**Perturbation Analysis of Parametric Resonance, Figure 4**  
Rotor with diskmass  $M$ , elastically mounted with axial ( $u$ ) and lateral directions

tion, frequency  $2\eta$ , the equations of motion become:

$$\begin{aligned}\ddot{x} + 2\alpha\dot{y} + (1 + 4\varepsilon\eta^2 \cos 2\eta t)x &= 0, \\ \ddot{y} - 2\alpha\dot{x} + (1 + 4\varepsilon\eta^2 \cos 2\eta t)y &= 0.\end{aligned}\quad (29)$$

System (29) constitutes a system of Mathieu-like equations, where we have neglected the effects of damping. Abbreviating  $P(t) = 4\eta^2 \cos 2\eta t$ , the corresponding Hamiltonian is:

$$\begin{aligned}H = \frac{1}{2}(1 + \alpha^2 + \varepsilon P(t))x^2 + \frac{1}{2}p_x^2 + \frac{1}{2}(1 + \alpha^2 + \varepsilon P(t))y^2 \\ + \frac{1}{2}p_y^2 + \alpha x p_y - \alpha y p_x,\end{aligned}$$

where  $p_x, p_y$  are the momenta. The natural frequencies of the unperturbed system (29),  $\varepsilon = 0$ , are  $\omega_1 = \sqrt{\alpha^2 + 1} + \alpha$  and  $\omega_2 = \sqrt{\alpha^2 + 1} - \alpha$ . By putting  $z = x + iy$ , system (29) can be written as:

$$\ddot{z} - 2\alpha i\dot{z} + (1 + 4\varepsilon\eta^2 \cos 2\eta t)z = 0. \quad (30)$$

Introducing the new variable:

$$v = e^{-i\alpha t} z, \quad (31)$$

and putting  $\eta t = \tau$ , we obtain:

$$v'' + \left( \frac{1 + \alpha^2}{\eta^2} + 4\varepsilon \cos 2\tau \right) v = 0, \quad (32)$$

where the prime denotes differentiation with respect to  $\tau$ . By writing down the real and imaginary parts of this equation, we get two identical Mathieu equations. We conclude that the trivial solution is stable for  $\varepsilon$  small enough, providing that  $\sqrt{1 + \alpha^2}$  is not close to  $n\eta$ , for some  $n = 1, 2, 3, \dots$ . The first-order interval of instability,  $n = 1$ , arises if:

$$\sqrt{1 + \alpha^2} \approx \eta. \quad (33)$$

If condition (33) is satisfied, the trivial solution of Eq. (32) is unstable. Therefore, the trivial solution of system (29) is also unstable. Note that this instability arises when:

$$\omega_1 + \omega_2 = 2\eta,$$

i.e. when the sum of the eigenfrequencies of the unperturbed system equals the excitation frequency  $2\eta$ . This is known as a sum resonance of first order. The domain of instability can be calculated as in Subsect. “Elementary The-

ory”; we find for the boundaries:

$$\eta_b = \sqrt{1 + \alpha^2} (1 \pm \varepsilon) + O(\varepsilon^2). \quad (34)$$

The second order interval of instability of Eq. (32),  $n = 2$ , arises when:

$$\sqrt{1 + \alpha^2} \approx 2\eta, \quad (35)$$

i.e.  $\omega_1 + \omega_2 \approx \eta$ . This is known as a sum resonance of second order. As above, we find the boundaries of the domains of instability:

$$2\eta = \sqrt{1 + \alpha^2} \left( 1 + \frac{1}{24}\varepsilon^2 \right) + O(\varepsilon^4), \quad (36)$$

$$2\eta = \sqrt{1 + \alpha^2} \left( 1 - \frac{5}{24}\varepsilon^2 \right) + O(\varepsilon^4).$$

Higher order combination resonances can be studied in the same way; the domains of instability in parameter space continue to narrow as  $n$  increases. It should be noted that the parameter  $\alpha$  is proportional to the rotation speed  $\Omega$  of the disk and to the ratio of the moments of inertia.

**Instability by Damping** We add small linear damping to system (29), with positive damping parameter  $\mu = 2\varepsilon\kappa$ . This leads to the equation:

$$\ddot{z} - 2\alpha i\dot{z} + (1 + 4\varepsilon\eta^2 \cos 2\eta t)z + 2\varepsilon\kappa\dot{z} = 0. \quad (37)$$

Because of the damping term, we can no longer reduce the complex Eq. (37) to two identical second order real equations, as we did in the previous section. In the sum resonance of the first order, we have  $\omega_1 + \omega_2 \approx 2\eta$  and the solution of the unperturbed ( $\varepsilon = 0$ ) equation can be written as:

$$z(t) = z_1 e^{i\omega_1 t} + z_2 e^{-i\omega_2 t}, \quad z_1, z_2 \in \mathbb{C}, \quad (38)$$

with  $\omega_1 = \sqrt{\alpha^2 + 1} + \alpha$ ,  $\omega_2 = \sqrt{\alpha^2 + 1} - \alpha$ . Applying variation of constants leads to equations for  $z_1$  and  $z_2$ :

$$\begin{aligned}\dot{z}_1 &= \frac{i\varepsilon}{\omega_1 + \omega_2} \left( 2\kappa (i\omega_1 z_1 - i\omega_2 z_2 e^{-i(\omega_1 + \omega_2)t}) \right. \\ &\quad \left. + 4\eta^2 \cos 2\eta t (z_1 + z_2 e^{-i(\omega_1 + \omega_2)t}) \right), \\ \dot{z}_2 &= \frac{-i\varepsilon}{\omega_1 + \omega_2} \left( 2\kappa (i\omega_1 z_1 e^{i(\omega_1 + \omega_2)t} - i\omega_2 z_2) \right. \\ &\quad \left. + 4\eta^2 \cos 2\eta t (z_1 e^{i(\omega_1 + \omega_2)t} + z_2) \right).\end{aligned}\quad (39)$$

To calculate the instability interval around the value  $\eta_0 = \frac{1}{2}(\omega_1 + \omega_2) = \sqrt{\alpha^2 + 1}$ , we apply perturbation the-

ory to find for the stability boundary:

$$\begin{aligned}\eta_b &= \sqrt{1 + \alpha^2} \left( 1 \pm \varepsilon \sqrt{1 + \alpha^2 - \frac{\kappa^2}{\eta_0^2}} + \dots \right), \\ &= \sqrt{1 + \alpha^2} \left( 1 \pm \sqrt{(1 + \alpha^2)\varepsilon^2 - \left(\frac{\mu}{\eta_0}\right)^2} + \dots \right).\end{aligned}\quad (40)$$

It follows that the domain of instability actually becomes *larger* when damping is introduced. The most unusual aspect of the above expression for the instability interval, however, is that there is a discontinuity at  $\kappa = 0$ . If  $\kappa \rightarrow 0$ , then the boundaries of the instability domain tend to the limits  $\eta_b \rightarrow \sqrt{1 + \alpha^2}(1 \pm \varepsilon\sqrt{1 + \alpha^2})$  which differs from the result we found when  $\kappa = 0$ :  $\eta_b = \sqrt{1 + \alpha^2}(1 \pm \varepsilon)$ .

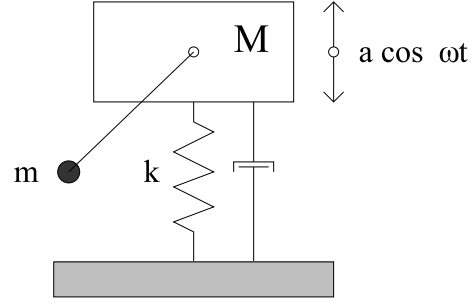
In mechanical terms, the broadening of the instability-domain is caused by the coupling between the two degrees of freedom of the rotor in lateral directions which arises in the presence of damping. Such phenomena are typical for gyroscopic systems and have been noted earlier in the literature; see [2,4] and [36]. The explanation of the discontinuity and its genericity in [17], see Subsect. “*Parametrically Forced Oscillators in Sum Resonance*”, is new. For hysteresis and phase-locking phenomena in this problem, the reader is referred to [33].

### Autoparametric Excitation

In [42], autoparametric systems are characterized as vibrating systems which consist of at least two consisting subsystems that are coupled. One is a Primary System that can be in normal mode vibration. In the instability (parameter) intervals of the normal mode solution in the full, coupled system, we have *autoparametric resonance*. The vibrations of the Primary System act as parametric excitation of the Secondary System which will no longer remain at rest. An example is presented in Fig. 5.

In actual engineering problems, we wish sometimes to diminish the vibration amplitudes of the Primary System; sometimes this is called ‘quenching of vibrations’. In other cases we have a coupled Secondary System which we would like to keep at rest. As an example we consider the following autoparametric system studied in [14]:

$$\begin{aligned}x'' + x + \varepsilon \left( k_1 x' + \sigma_1 x + a \cos 2\tau x + \frac{4}{3} x^3 + c_1 y^2 x \right) &= 0 \\ y'' + y + \varepsilon \left( k_2 y' + \sigma_2 y + c_2 x^2 y + \frac{4}{3} y^3 \right) &= 0\end{aligned}\quad (41)$$



**Perturbation Analysis of Parametric Resonance, Figure 5**

**Two coupled oscillators with vertical oscillations as Primary System and parametric excitation of the coupled pendulum (Secondary System)**

where  $\sigma_1$  and  $\sigma_2$  are the detunings from the 1 : 1-resonance of the oscillators. In this system,  $y(t) = \dot{y}(t) = 0$  corresponds with a normal mode of the  $x$ -oscillator.

The system (41) is invariant under  $(x, y) \rightarrow (x, -y)$ ,  $(x, y) \rightarrow (-x, y)$ , and  $(x, y) \rightarrow (-x, -y)$ . Using the method of averaging as a normalization procedure we investigate the stability of solutions of system (41). To give an explicit example we follow [14] in more detail. Introduce the usual variation of constants transformation:

$$x = u_1 \cos \tau + v_1 \sin \tau ; \quad x' = -u_1 \sin \tau + v_1 \cos \tau \quad (42)$$

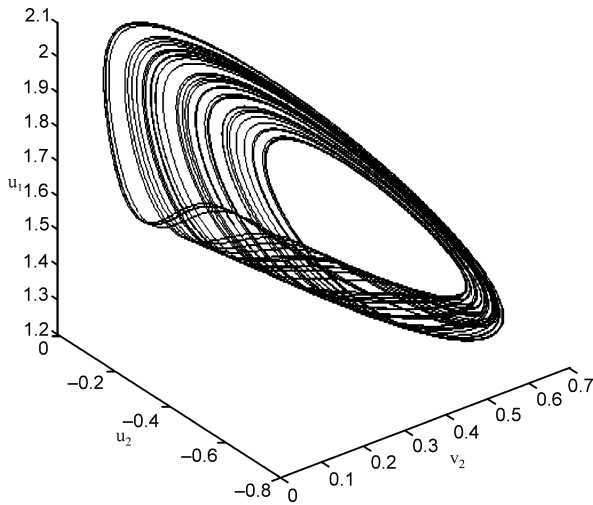
$$y = u_2 \cos \tau + v_2 \sin \tau ; \quad y' = -u_2 \sin \tau + v_2 \cos \tau \quad (43)$$

After rescaling  $\tau = \frac{\varepsilon}{2} \tilde{\tau}$  the averaged system of (41) becomes:

$$\begin{aligned}u_1' &= -k_1 u_1 + \left( \sigma_1 - \frac{1}{2} a \right) v_1 + v_1 (u_1^2 + v_1^2) \\ &\quad + \frac{1}{4} c_1 u_2^2 v_1 + \frac{3}{4} c_1 v_2^2 v_1 + \frac{1}{2} c_1 u_2 v_2 u_1 \\ v_1' &= -k_1 v_1 - \left( \sigma_1 + \frac{1}{2} a \right) u_1 - u_1 (u_1^2 + v_1^2) \\ &\quad - \frac{3}{4} c_1 u_2^2 u_1 - \frac{1}{4} c_1 v_2^2 u_1 - \frac{1}{2} c_1 u_2 v_2 v_1 \\ u_2' &= -k_2 u_2 + \sigma_2 v_2 + v_2 (u_2^2 + v_2^2) + \frac{1}{4} c_2 u_1^2 v_2 \\ &\quad + \frac{3}{4} c_2 v_1^2 v_2 + \frac{1}{2} c_2 u_1 v_1 u_2 \\ v_2' &= -k_2 v_2 - \sigma_2 u_2 - u_2 (u_2^2 + v_2^2) - \frac{3}{4} c_2 u_1^2 u_2 \\ &\quad - \frac{1}{4} c_2 v_1^2 u_2 - \frac{1}{2} c_2 u_1 v_1 v_2.\end{aligned}\quad (44)$$

This system is analyzed for critical points, periodic and quasi-periodic solutions, producing existence and stability diagrams in parameter space. The system also con-





**Perturbation Analysis of Parametric Resonance, Figure 6**  
The strange attractor of the averaged system (44). The phase portraits in the  $(u_2, v_2, u_1)$ -space for  $c_2 < 0$  at the value  $\sigma_2 = 5.3$ . The Kaplan–Yorke dimension for  $\sigma_2 = 5.3$  is 2.29

tains a sequence of period-doubling bifurcations leading to chaotic solutions, see Fig. 6.

To prove the presence of chaos involves an application of higher dimensional Melnikov theory developed in [47]. A rather technical analysis in [14] shows the existence of a Šilnikov orbit in the averaged equation, which implies chaotic dynamics, also for the original system.

### Future Directions

Ongoing research in dynamical systems includes nonlinear systems with parametric resonance, but there are a number of special features as these systems are non-autonomous. This complicates the dynamics from the outset. For instance a two degrees of freedom system with parametric resonance involves at least three frequencies, producing many possible resonances. The analysis of such higher dimensional systems with many more combination resonances, has begun recently, producing interesting limit sets and invariant manifolds. Also the analysis of PDEs with periodic coefficients will play a part in the near future. These lines of research are of great interest.

In the conservative case, the association with Hamiltonian systems, KAM theory etc. gives a natural approach. This has already produced important results. In real-life modeling, there will always be dissipation and it is important to include this effect. Preliminary results suggest that the impact of damping on for instance quasi-periodic systems, is quite dramatic. This certainly merits more research.

Finally, applications are needed to solve actual problems and to inspire new, theoretical research.

### Acknowledgment

A number of improvements and clarifications were suggested by the editor, Giuseppe Gaeta. Additional references were obtained from Henk Broer and Fadi Dohnal.

### Bibliography

#### Primary Literature

1. Arnold VI (1983) Geometrical methods in the theory of ordinary differential equations. Springer, New York
2. Banichuk NV, Bratus AS, Myshkis AD (1989) Stabilizing and destabilizing effects in nonconservative systems. *PMM USSR* 53(2):158–164
3. Bogoliubov NN, Mitropolskii Yu A (1961) Asymptotic methods in the theory of nonlinear oscillations. Gordon and Breach, New York
4. Bolotin VV (1963) Non-conservative problems of the theory of elastic stability. Pergamon Press, Oxford
5. Broer HW, Vegter G (1992) Bifurcational aspects of parametric resonance, vol 1. In: Jones CKRT, Kirchgraber U, Walther HO (eds) Expositions in dynamical systems. Springer, Berlin, pp 1–51
6. Broer HW, Levi M (1995) Geometrical aspects of stability theory for Hill's equation. *Arch Rat Mech Anal* 131:225–240
7. Broer HW, Simó C (1998) Hill's equation with quasi-periodic forcing: resonance tongues, instability pockets and global phenomena. *Bol Soc Brasil Mat* 29:253–293
8. Broer HW, Hoveijn I, Van Noort M (1998) A reversible bifurcation analysis of the inverted pendulum. *Physica D* 112:50–63
9. Broer HW, Hoveijn I, Van Noort M, Vegter G (1999) The inverted pendulum: a singularity theory approach. *J Diff Eqs* 157:120–149
10. Broer HW, Simó C (2000) Resonance tongues in Hill's equations: a geometric approach. *J Differ Equ* 166:290–327
11. Broer HW, Puig J, Simó C (2003) Resonance tongues and instability pockets in the quasi-periodic Hill-Schrödinger equation. *Commun Math Phys* 241:467–503
12. Broer HW, Hoveijn I, Van Noort M, Simó C, Vegter G (2005) The parametrically forced pendulum: a case study in  $1\frac{1}{2}$  degree of freedom. *J Dyn Diff Equ* 16:897–947
13. Cicogna G, Gaeta G (1999) Symmetry and perturbation theory in nonlinear dynamics. *Lecture Notes Physics*, vol 57. Springer, Berlin
14. Fatimah S, Ruijgrok M (2002) Bifurcation in an autoparametric system in 1:1 internal resonance with parametric excitation. *Int J Non-Linear Mech* 37:297–308
15. Golubitsky M, Schaeffer D (1985) Singularities and groups in bifurcation theory. Springer, New York
16. Hale J (1963) Oscillation in nonlinear systems. McGraw-Hill, New York, 1963; Dover, New York, 1992
17. Hoveijn I, Ruijgrok M (1995) The stability of parametrically forced coupled oscillators in sum resonance. *ZAMP* 46:383–392
18. Iooss G, Adelmeyer M (1992) Topics in bifurcation theory. World Scientific, Singapore

19. Kuznetsov Yu A (2004) Elements of applied bifurcation theory, 3rd edn. Springer, New York
20. Krupa M (1997) Robust heteroclinic cycles. *J Nonlinear Sci* 7:129–176
21. Len JL, Rand RH (1988) Lie transforms applied to a non-linear parametric excitation problem. *Int J Non-linear Mech* 23:297–313
22. Levy DM, Keller JB (1963) Instability intervals of Hill's equation. *Comm Pure Appl Math* 16:469–476
23. Magnus W, Winkler S (1966) Hill's equation. Interscience-John Wiley, New York
24. McLaughlin JB (1981) Period-doubling bifurcations and chaotic motion for a parametrically forced pendulum. *J Stat Phys* 24:375–388
25. Ng L, Rand RH (2002) Bifurcations in a Mathieu equation with cubic nonlinearities. *Chaos Solitons Fractals* 14:173–181
26. Ng L, Rand RH (2003) Nonlinear effects on coexistence phenomenon in parametric excitation. *Nonlinear Dyn* 31:73–89
27. Pikovsky AS, Feudel U (1995) Characterizing strange non-chaotic attractors. *Chaos* 5:253–260
28. Ramani DV, Keith WL, Rand RH (2004) Perturbation solution for secondary bifurcation in the quadratically-damped Mathieu equation. *Int J Non-linear Mech* 39:491–502
29. Recktenwald G, Rand RH (2005) Coexistence phenomenon in autoparametric excitation of two degree of freedom systems. *Int J Non-linear Mech* 40:1160–1170
30. Roseau M (1966) Vibrations nonlinéaires et théorie de la stabilité. Springer, Berlin
31. Ruijgrok M (1995) Studies in parametric and autoparametric resonance. Thesis, Utrecht University, Utrecht
32. Ruijgrok M, Verhulst F (1996) Parametric and autoparametric resonance. *Prog Nonlinear Differ Equ Their Appl* 19:279–298
33. Ruijgrok M, Tondl A, Verhulst F (1993) Resonance in a rigid rotor with elastic support. *ZAMM* 73:255–263
34. Sanders JA, Verhulst F, Murdock J (2007) Averaging methods in nonlinear dynamical systems, rev edn. *Appl Math Sci*, vol 59. Springer, New York
35. Seyranian AP (2001) Resonance domains for the Hill equation with allowance for damping. *Phys Dokl* 46:41–44
36. Seyranian AP, Mailybaev AA (2003) Multiparameter stability theory with mechanical applications. Series A, vol 13. World Scientific, Singapore
37. Seyranian AA, Seyranian AP (2006) The stability of an inverted pendulum with a vibrating suspension point. *J Appl Math Mech* 70:754–761
38. Stoker JJ (1950) Nonlinear vibrations in mechanical and electrical systems. Interscience, New York, 1950; Wiley, New York, 1992
39. Strutt MJO (1932) Lamé-sche, Mathieu-sche und verwandte Funktionen. Springer, Berlin
40. Szeplinska-Stupnicka W (1990) The behaviour of nonlinear vibrating systems, vol 2. Kluwer, Dordrecht
41. Tondl A (1991) Quenching of self-excited vibrations. Elsevier, Amsterdam
42. Tondl A, Ruijgrok M, Verhulst F, Nabergoj R (2000) Autoparametric resonance in mechanical systems. Cambridge University Press, New York
43. Van der Pol B, Strutt MJO (1928) On the stability of the solutions of Mathieu's equation. *Phil Mag Lond Edinb Dublin* 7(5):18–38
44. Verhulst F (1996) Nonlinear differential equations and dynamical systems. Springer, New York
45. Verhulst F (2005) Invariant manifolds in dissipative dynamical systems. *Acta Appl Math* 87:229–244
46. Verhulst F (2005) Methods and applications of singular perturbations. Springer, New York
47. Wiggins S (1988) Global Bifurcation and Chaos. *Appl Math Sci*, vol 73. Springer, New York
48. Yakubovich VA, Starzhinskii VM (1975) Linear differential equations with periodic coefficients, vols 1 and 2. Wiley, New York
49. Zounes RS, Rand RH (1998) Transition curves for the quasi-periodic Mathieu equation. *SIAM J Appl Math* 58:1094–1115
50. Zounes RS, Rand RH (2002) Global behavior of a nonlinear quasi-periodic Mathieu equation. *Nonlinear Dyn* 27:87–105

### Books and Reviews

- Arnold VI (1977) Loss of stability of self-oscillation close to resonance and versal deformation of equivariant vector fields. *Funct Anal Appl* 11:85–92
- Arscott FM (1964) Periodic differential equations. MacMillan, New York
- Cartmell M (1990) Introduction to linear, parametric and nonlinear vibrations. Chapman and Hall, London
- Dohnal F (2005) Damping of mechanical vibrations by parametric excitation. Ph D thesis, Vienna University of Technology
- Dohnal F, Verhulst F (2008) Averaging in vibration suppression by parametric stiffness excitation. *Nonlinear Dyn* (accepted for publication)
- Ecker H (2005) Suppression of self-excited vibrations in mechanical systems by parametric stiffness excitation. *Fortschrittsberichte Simulation Bd 11*. Argesim/Asim Verlag, Vienna
- Fatimah S (2002) Bifurcations in dynamical systems with parametric excitation. Thesis, University of Utrecht
- Hale J (1969) Ordinary differential equations. Wiley, New York
- Kirillov ON (2007) Gyroscopic stabilization in the presence of nonconservative forces. *Dokl Math* 76:780–785; *Orig Russian*: (2007) *Dokl Ak Nauk* 416:451–456
- Meixner J, Schäfer FW (1954) Mathiesche Funktionen und Sphäroidfunktionen. Springer, Berlin
- Moon FC (1987) Chaotic vibrations: an introduction for applied scientists and engineers. Wiley, New York
- Nayfeh AH, Mook DT (1979) Nonlinear Oscillations. Wiley Interscience, New York
- Schmidt G (1975) Parametererregte Schwingungen. VEB Deutscher Verlag der Wissenschaften, Berlin
- Schmidt G, Tondl A (1986) Non-linear vibrations. Akademie-Verlag, Berlin
- Tondl A (1978) On the interaction between self-excited and parametric vibrations. In: Monographs and Memoranda, vol 25. National Res Inst Běchovice, Prague
- Tondl A (1991) On the stability of a rotor system. *Acta Technica CSAV* 36:331–338
- Tondl A (2003) Combination resonances and anti-resonances in systems parametrically excited by harmonic variation of linear damping coefficients. *Acta Technica CSAV* 48:239–248
- Van der Burgh AHP, Hartono (2004) Rain-wind induced vibrations of a simple oscillator. *Int J Non-Linear Mech* 39:93–100
- Van der Burgh AHP, Hartono, Abramian AK (2006) A new model for the study of rain-wind-induced vibrations of a simple oscillator. *Int J Non-Linear Mech* 41:345–358

- Weinstein A, Keller JB (1985) Hill's equation with a large potential. *SIAM J Appl Math* 45:200–214
- Weinstein A, Keller JB (1987) Asymptotic behaviour of stability regions for Hill's equation. *SIAM J Appl Math* 47:941–958
- Wiggins S (1990) Introduction to applied nonlinear dynamical systems and chaos. Springer, New York

## Perturbation of Equilibria in the Mathematical Theory of Evolution

ANGEL SÁNCHEZ<sup>1,2</sup>

<sup>1</sup> Grupo Interdisciplinar de Sistemas Complejos (GISC), Departamento de Matemáticas, Universidad Carlos III de Madrid, Madrid, Spain

<sup>2</sup> Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Zaragoza, Spain

### Article Outline

Glossary

Definition of the Subject

Introduction

Evolution on a Fitness Landscape

Stability of Equilibria on a Fitness Landscape

Perturbation of Equilibria on a Fitness Landscape

Frequency Dependent Fitness: Game Theory

Equilibria in Evolutionary Game Theory

Perturbations of Equilibria

in Evolutionary Game Theory

Future Directions

Bibliography

### Glossary

**Evolutionarily stable equilibria (ESS)** An ESS is a set of frequencies of different types of individuals in a population that can not be invaded by the evolution of a single mutant. It is the evolutionary counterpart of a Nash equilibrium.

**Fitness landscape** A metaphorical description of fitness as a function of individual's genotypes or phenotypes in terms of a multivariable function that does not depend on any external influence.

**Genetic locus** The position of a gene on a chromosome. The different variants of the gene that can be found at the same locus are called alleles.

**Nash equilibrium** In classical game theory, a Nash equilibrium is a set of strategies, one for each player of the game, such that none of them can improve her benefits by unilateral changes of strategy.

**Scale free network** A graph or network such that the degrees of the nodes are taken from a power-law distribution. As a consequence, there is not a typical degree in the graph, i. e., there are no typical scales.

**Small-world network** A graph or network of  $N$  nodes such that the mean distance between nodes scales as  $\log N$ . It corresponds to the well-known “six degrees of separation” phenomenon.

### Definition of the Subject

The importance of evolution can hardly be overstated. As the Jesuit priest Pierre Teilhard de Chardin put it,

Evolution is a general postulate to which all theories, all hypotheses, all systems must hence forward bow and which they must satisfy in order to be thinkable and true. Evolution is a light which illuminates all facts, a trajectory which all lines of thought must follow – this is what evolution is.

Darwin's evolution theory is based on three fundamental principles: reproduction, mutation and selection, which describe how populations change over time and how new forms evolve out of old ones. Starting with W. F. R. Weldon, whom at the beginning of the 20th century realized that “the problem of animal evolution is essentially a statistical problem”, and blooming in the 30's with Fisher, Haldane and Wright, numerous mathematical descriptions of the resulting evolutionary dynamics have been proposed, developed and studied. Deeply engraved in these frameworks are the mathematical concepts of equilibrium and stability, as descriptions of the observed population compositions and their lifetimes. Many results have been obtained regarding the stability of equilibria of evolutionary dynamics in idealized circumstances, such as infinite populations or global interactions. In the evolutionary context, stability is peculiar, in the sense that it is entangled with collective effects arising from the interaction of individuals. Therefore, perturbations of the idealized mathematical framework representing more realistic situations are of crucial importance to understand stability of equilibria.

### Introduction

The idea of evolution is a simple one: Descent with modification acted upon by natural selection. Descent with modification means that we consider a population of replicators, entities capable of reproducing themselves, in which reproduction is not exact and allows for small differences between parents and offspring. Natural selection means

that different entities reproduce in different quantities because their abilities are also different: some are more resistant to external factors, some need less resources, some simply reproduce more... While in biology these replicators are, of course, living beings, the basic ingredients of evolution have by now transcended the realm of biology into the kingdom of objects such as computer codes (thus giving rise, e. g., to genetic algorithms).

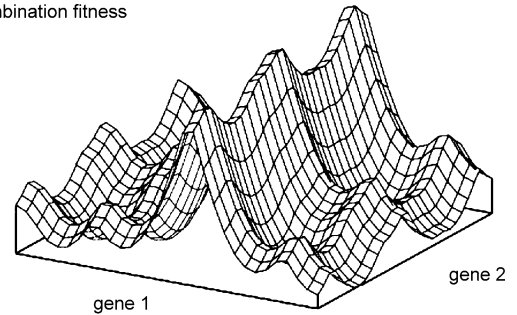
The process of evolution is often described in short as “the survival of the fittest”, meaning that those replicators that succeed and thrive are more fit than those which progressively disappear. This statement is not very appropriate, because evolution does not imply making organisms or entities more fit; it is simply a consequence of differential reproduction in the face of selection pressure. On the other hand, it leads to a tautology: The question of which are the fittest organisms is answered by saying that they are those that survive. It is then clear that a correct use of the concept of fitness is at the crux of any attempt to formalize mathematically evolution theory.

In this article we are going to discuss two manners to deal mathematically with the concept of fitness. The first and simplest one is to resort to a *fitness landscape*, whose basic feature is that the fitness of a given individual depends only on the individual’s characteristics and not on external factors. We will present this approach in Sects. “[Evolution on a Fitness Landscape](#)” and “[Stability of Equilibria on a Fitness Landscape](#)” below, to subsequently discuss the effect of perturbations on the equilibria described by this picture in Sect. “[Perturbation of Equilibria on a Fitness Landscape](#)”. To go beyond the fitness landscape picture one has to introduce frequency-dependent selection, i. e., to remove the independence of the fitness from external influences. In Sects. “[Frequency Dependent Fitness: Game Theory](#)” and “[Equilibria in Evolutionary Game Theory](#)” we consider the *evolutionary game theory* approach to this way to model evolution and, as before, analyze the perturbations of its equilibria in Sec “[Perturbations of Equilibria in Evolutionary Game Theory](#)”. Sect. “[Future Directions](#)” summarizes the questions that remain open in this field.

### Evolution on a Fitness Landscape

The metaphor of evolution on a “fitness landscape” reaches back at least to [26]: Drawing on the connection between fitness and adaptation, fitness is defined as the expected number of offspring of a given individual that reach adulthood, and thus represents a measure of its adaptation to the environment. In this context, fitness landscapes are used to visualize the relationship between genotypes

gene combination fitness



**Perturbation of Equilibria in the Mathematical Theory of Evolution, Figure 1**

**Sketch of a fitness landscape**

(or phenotypes) and reproductive success. It is assumed that every genotype has a well defined fitness, in the sense above, and that this fitness is the “height” of the landscape. Genotypes which are very similar are said to be “close” to each other, while those that are very different are “far” from each other. The two concepts of height and distance are sufficient to form the concept of a “landscape”. The set of all possible genotypes, their degree of similarity, and their related fitness values is then called a fitness landscape.

Fitness landscapes are often conceived of as ranges of mountains. There exist local peaks (points from which all paths are downhill, i. e. to lower fitness) and valleys (regions from which most paths lead uphill). A fitness landscape with many local peaks surrounded by deep valleys is called rugged. If all genotypes have the same replication rate, on the other hand, a fitness landscape is said to be flat. A sketch of such a fitness landscape, showing the dependence of the fitness on two different “characteristics” or “genes”, is shown in Fig. 1. Of course, the true fitness landscape would need a highly multidimensional space for its representation, as it would depend on all the characteristics of the organism, even those it still does not show. Therefore, the sketch is an extreme oversimplification, only to suggest the structure of a rugged fitness landscape.

Given the immense complexity of the genotype-fitness mapping, theoretical models have to make a variety of simplifying assumptions. Most models in biological literature focus on the effect of one or a few genetic loci on the fitness of individuals in a population, assuming that each of the considered loci can be occupied by a limited number of different alleles that have different effects on the fitness, and that the rest of the genome is part of the invariant environment. This approximation, the first attempt to obtain analytical results for changes in the gene pool of a population under the influence of inheritance,

selection and mutation is the pioneering work of Fisher, Haldane and Wright, who founded the field of population genetics. Their method of randomly drawing the genes of the daughter population from the pool of parent genes, with weights proportional to the fitness, proved to be very successful at calculating the evolution of allele frequencies from one generation to the next, or the chances of a new mutation to spread through a population, even taking into account various patterns of mating, dominance effects, nonlinear effects between different genes, etc. Population genetics has since then developed into a mature field with a sophisticated mathematical apparatus, and with wide-ranging applications.

### Stability of Equilibria on a Fitness Landscape

The simplified pictures we have just described lead to a description in terms of dynamical systems, and therefore the stability of its equilibria can be studied by means of standard techniques. In principle, one can envisage the evolution of a population on a fitness landscape in the usual frame of particle dynamics on a potential. Every individual is a point in the space of phenotypes or genotypes, and evolves towards the maxima of the fitness; in the potential picture, the potential is given by minus the fitness. Furthermore, the dynamics is overdamped, i. e., there are no oscillations around the equilibria. Maxima of the fitness are therefore the equilibria of the evolutionary process. That this is so is a consequence of Fisher's theorem [3], whose original derivation is very general but quite complicated. Following [2] and [4], we prefer here to present two simpler situations: an asexual population, and a sexually reproducing population where the fitness is determined by a single gene with two alleles.

For an asexually reproducing population, the derivation of Fisher's theorem is straightforward: Let  $y_i$  be the number of genes  $i$  in the population, and  $y$  the total number of genes; then  $p_i \equiv y_i/y$  is the frequency of genotype  $i$  in the population. If  $W_i$  is gene  $i$ 's fitness, sticking to the interpretation of fitness in terms of offspring, the number of individuals carrying gene  $i$  in the next generation is  $W_i y_i$  and, subsequently, the change in the frequency  $p_i$  from one generation to the next is

$$\Delta p_i = p_i(W_i - \bar{W})/\bar{W},$$

leading to a change in mean fitness

$$\frac{\Delta \bar{W}}{\bar{W}} = \frac{\sum_i W_i \Delta p_i}{\bar{W}} = \frac{\sum_i p_i (W_i^2 - \bar{W}^2)}{\bar{W}^2},$$

which is proportional to the genetic variance in fitness. If the fitness changes from one generation to the next are small, this becomes an equation which states that the rate of change in fitness is identical to the genetic variance in fitness.

When reproduction is sexual, we note that  $p_i$  is the frequency of gene  $i$ . Assuming for simplicity that only two alleles are possible at the genetic locus of interest, the fitness of type 1 is  $w_{11}p_1 + w_{12}p_2$ , where  $w_{ij}$  is the fitness of an individual carrying alleles  $i$  and  $j$ , and hence the number of 1 genes will be  $y_1(w_{11}p_1 + w_{12}p_2)$ . We then have the differential equation

$$\dot{y}_1 = y_1(w_{11}p_1 + w_{12}p_2). \quad (1)$$

It is enough then to differentiate the identity  $\ln p_1 = \ln y_1 - \ln y$  and use a little algebra to show that  $y_1$  obeys the replicator equation:

$$\dot{p}_1 = p_1(w_1 - \bar{w}) \quad (2)$$

Fisher's theorem then states that fitness increases along trajectories of this equation: Indeed, by noting that the average fitness is

$$\bar{w} = \sum_{i,j=1,2} w_{ij}p_i p_j, \quad (3)$$

differentiating and using the replicator equation we arrive at the final result

$$\dot{\bar{w}} = 2 \sum_i p_i (w_i - \bar{w})^2. \quad (4)$$

Fisher's theorem thus means that an evolving population will typically climb uphill in the fitness landscape, by a series of small genetic changes, until a local optimum is reached. This is due to the fact that the average fitness of the population always increases, as we have just shown; hence the analogy with overdamped dynamics on a (inverted) potential function. Furthermore, because of this result, the population remains there, at the equilibrium point, because it cannot reduce its fitness. We then realize that all equilibria in a fitness landscape within the interpretation of fitness as the reproductive success are stable.

### Perturbation of Equilibria on a Fitness Landscape

In order to understand the possible breakdowns of stability in the fitness landscape picture, one has to look carefully at the hypothesis of Fisher's theorem. We have not stated it in a formal manner, hence it is important to summarize here the main ones:



- Population is infinite.
- There are no mutations (i. e., the only source of every gene or species is reproduction).
- There is only one population (i. e., there are no population fluxes or migrations between separate groups).
- Fitness depends only on the individual's genotype and not on the other individuals.

It is then clear that, even if it is mathematically true, the applicability of Fisher's theorem is a completely different story, and as a consequence, the conclusion that maxima of the fitness landscape are stable may be wrong when discussing real systems. A detailed discussion of all these issues can be found in [2], and we refer the reader to her paper for a thorough discussion of all these factors. For our present purposes, namely to show that these perturbations can change the stability of the equilibria, it will suffice to present a few ideas about the case of finite population size. Afterwards, the rest of the paper will proceed along the idea of fitness depending on other individuals, giving up the paradigm of a fixed fitness landscape.

The subject of finite size populations is the subject of fluctuations and its main consequences, genetic drift and stochastic escape. Regarding the first concept, as compared to natural selection, i. e., to the tendency of beneficial alleles to become more common over time (and detrimental ones less common), genetic drift is the fundamental tendency of any allele to vary randomly in frequency over time due to statistical variation alone, so long as it does not comprise all or none of the distribution. In other words, even when individuals face the same odds, they will differ in their success. A rare succession of chance events can thus bring a trait to predominance, causing a population or species to evolve (in fact, this idea is at the core of the neutral theory of evolution, first proposed by [10]). On the other hand, stochastic escape refers to the situation in which a population of individuals placed at a maximum of the fitness landscape may leave this maximum due to fluctuations. Obviously, both genetic drift and stochastic escape affect the stability of the maxima as predicted by Fisher's theorem.

One consequence of finite population sizes and fluctuations in the composition of a population is that genes get lost from the gene pool. If there is no new genetic input through mutation or migration, the genetic variability within a population decreases with time. After sufficiently many generations, all individuals will carry the same allele of a given gene. This allele is said to have become fixed. In the absence of selection, the probability that a given allele will become fixed is proportional to the number of copies in the initial population. Thus, if a new mutant arises that

has no selective advantage or disadvantage, this mutant will spread through the entire population with a probability  $1/M$ ,  $M$  being the population size. If the individuals of the population are diploid, each carries two sets of genes, and  $M$  must be taken as the number of sets of genes, i. e., as twice the population size. On the other hand, it can be shown [2] that the probability that a mutant that conveys a small fitness increase by a factor  $1 + s$  has as probability of the order  $s$  to spread through a population. In populations of sizes much smaller than  $1/s$ , this selective advantage is not felt, because mutations that carry no advantage become fixed at a similar rate. In the same manner, a mutation that decreases the fitness of its carrier by a factor  $1 - s$ , is not felt in a population much smaller than  $1/s$ . An interesting consequence of these results is that the rate of neutral (or effectively neutral) substitutions is independent of the population size. The reason is that the probability that a new mutant is generated in the population is proportional to  $M$ , while its probability of becoming fixed is  $1/M$ .

### Frequency Dependent Fitness: Game Theory

In the preceding sections we have considered the case when the fitness depends on the genotype, but is independent of the composition of the population, i. e., the presence of individuals of the same genotype or of other genotypes does not change the fitness of the focal one. This assumption, that allows for an intuitive picture in terms of a fitness landscape, is clearly an over-simplification, as was already mentioned above. For instance, consider an homogeneous population in a closed environment. The population will grow at a pace given by the fitness of its individuals until it eventually exhausts the available resources or even physically fills the environment. Therefore, even if the individuals are all equal, their fitness will not be the same if there are only a few of them or if there are very many. Another trivial example is the effect on the fitness of the presence or absence of predators of the species of interest; clearly, predators will reduce the fitness (understood as above in a reproductive sense) of their prey.

Therefore, individuals will evolve subject not only to external influences but also to their mutual competition, both intra-specific and inter-specific. This leads us to consider frequency-dependent selection, which can be described by very many, different theoretical approaches. These include game theory as well as discrete and continuous genetic models, and the concepts of kin selection, group selection, and sexual selection. Among the possible dynamical patterns arising, there are single fixed points, lines of fixed points, runaway, limit cycles, and chaos. A re-

view of all these descriptions, whose use to model evolution depends on the specific issues one is interested in, is clearly far beyond the scope of this article and, hence, we have focused on evolutionary game theory as a particularly suited case study to show the effects of perturbations on equilibria.

Before going into the study of evolutionary game theory, we need to summarize briefly a few key concepts about its originating theory, namely (classical) game theory. Pioneered in the early XIX century by the economist Cournot, game theory was introduced by the brilliant, multi-faceted mathematician John von Neumann in 1928, and it was first presented as a specific subject in von Neumann's book (with Oskar Morgenstern) *Theory of Games and Economic Behavior* in 1944. Since then, game theory has been used to model strategic situations, i. e., situations in which actors or agents follow different strategies (meaning that they choose among different possible actions or behaviors) to maximize their benefit, usually referred to as *payoff*. These arise in very many different contexts, from biology and psychology to philosophy through politics, economics or computer science.

The central concept of game theory is the *Nash equilibrium* [14], introduced by the mathematician John Nash in 1955, awarded with a Nobel Prize in Economics almost forty years later for this work. A set of strategies, one for each participant in the game, is a Nash equilibrium if every strategy is the best response (in terms of maximizing the player's payoff) to the subset of the strategies of the rest of the players. In this case, if all players use strategies belonging to a Nash equilibrium, none of them will have any incentive to change her behavior. In this situation we indeed have the equivalent of the traditional concept of equilibrium in dynamical systems: players keep playing the same strategy as, given the behavior of the others, they follow the optimal strategy (note that this does not mean the strategy is optimal in absolute terms: it is only optimal in view of the actions of the rest).

## Equilibria in Evolutionary Game Theory

In the seventies, game theory, which as proposed by von Neumann and Nash was to be used to understand economic behavior, entered the realm of biology through the pioneering work of John Maynard-Smith and George Price [12], who introduced the evolutionary version of the theory. The key contribution of their work was a new interpretation of the general framework of game theory in terms of populations instead of individual players. While traditional game theoretical players behaved following some strategy and could change it to improve

their performance, in the picture of Maynard-Smith and Price individuals had a fixed strategy, determined by their genotype, and different strategies were represented by sub-populations of individuals. In this representation, changes of strategies correspond to the replacement of the individuals by their offspring, possibly with mutations. Payoffs obtained by individuals in the game are accordingly understood as fitness, the reproductive rate that governs how the replacement occurs.

There is a large degree of arbitrariness as to the evolutionary dynamics of the populations. All we have said so far is that fitness, obtained through the game, determines the composition of the population at the next time step (or instant, if we think of continuous time). Probably the most popular choice (but by no means the only one, see [18] for different evolutionary proposals and their relationships, see also [9] for other dynamics) is to use the replicator equation we have previously found to describe the evolution of the frequency  $y_i$  of strategists of type  $i$ :

$$\dot{y}_i = y_i(w_i - \bar{w}). \quad (5)$$

It is important to note that the steps leading to the derivation of this equation are the same as above, and therefore for it to be applicable in principle one must keep in mind the same hypotheses. The difference is that now the fitness is not a constant but rather it is determined by a game, which enters the equation in the following way.

Let us call  $\mathbf{A}$  the payoff matrix of the game (for simplicity, we will consider only symmetric games), whose entries  $a_{ij}$  are the payoffs to an individual using strategy  $i$  facing another using strategy  $j$ . Assuming the frequencies  $y_i(t)$  are differentiable functions, if individuals meet randomly and then engage in the game, and this takes place very many (infinite) times, then  $(\mathbf{A}\mathbf{y})_i$  is the expected payoff for type  $i$  individuals in a population described by the vector  $\mathbf{y}$ , whose components are the frequencies of each type. By the same token, the average payoff in the population is  $\bar{w} = \mathbf{y}^T \mathbf{A} \mathbf{y}$ , so substituting in (5) we are left with

$$\dot{y}_i = y_i[(\mathbf{A}\mathbf{y})_i - \mathbf{y}^T \mathbf{A} \mathbf{y}], \quad (6)$$

where we now see explicitly how the game affects the evolution. Nevertheless, it is also clear that this rule is arbitrary, and there are many other options one can use to postulate how the population evolves. We will come back to this issue when considering perturbations of the equilibria.

If Nash equilibrium is the key concept in game theory, evolutionarily stable strategy is the relevant one in its evolutionary counterpart. [11] defined evolutionarily stable strategy (ESS) as a strategy such that, if every individual

in the population uses it, no other (mutant) strategy could invade the population by natural selection. It is trivial to show that, in terms of the payoffs of the game, for strategy  $i$  to be an ESS, one of the following two conditions must hold:

$$a_{ii} > a_{ij} \quad \forall j \neq i, \quad \text{or} \quad (7)$$

$$a_{ii} = a_{ij} \text{ for some } j \quad \text{and} \quad a_{ij} > a_{jj}. \quad (8)$$

If the first condition is fulfilled, we speak of *strict* ESS. It is important to realize that this concept is absolutely general and, in particular, it does not depend on the evolutionary dynamics of choice (in so far as it favors the strategies that receive the best payoffs).

Of course, the two concepts, Nash equilibrium and ESS, are related. This is in fact one of the reasons why evolutionary game theory ended up appealing to the economists, who faced the question as to how individuals ever get to play the Nash equilibrium strategies: They now had a dynamical way that might precisely describe that process and, furthermore, to decide which Nash equilibrium was selected if there were more than one. To show the connection, one must decide on a dynamical rule, for which we will stay within the framework of the replicator dynamics. For this specific evolutionary dynamics, it can be rigorously shown that (see, e. g., [7,9])

1. if  $y_0$  is a Nash equilibrium, it is a rest point (a zero of the rhs of (5));
2. if  $y_0$  is a strict Nash equilibrium, it is asymptotically stable;
3. if  $y_0$  is a rest point and is the limit of an interior orbit for  $t \rightarrow \infty$ , then it is a Nash equilibrium; and
4. if  $y_0$  is a stable rest point, it is a Nash equilibrium.

This means that there indeed is a relationship between Nash equilibria and ESS, but more subtle that could appear at first. Probably, the most important non trivial aspect of this result is that not all ESS are Nash equilibria, as stability is required in addition.

### Perturbations of Equilibria in Evolutionary Game Theory

The evolutionary viewpoint on game theory allows to study Nash equilibria/ESS within the standard framework of dynamical systems theory, by using the concepts of stability, asymptotical stability, global stability and related notions. In fact, one can do more than that: the problem of invasion by a mutant, the biological basis of the ESS concept of Maynard-Smith, can always be formulated in

terms of a dynamical coupling of the mutant and the incumbent species and hence studied in terms of the stability of a rest point of a dynamical system. In principle, the same idea can be generalized to simultaneous invasion by more than one mutant and, although the problem may be technically much more difficult, the basic procedure remains the same.

As we did with the fitness landscape concept, when considering perturbations of equilibria, our interest goes beyond this traditional stability ideas, and once again, we need to focus on the deviations from the framework that allows to derive the replicator equation. There are a number of such deviations. The simplest ones are the inclusion of mutations or migrations, leading to the so-called replicator-mutator equation [18], that can be subsequently studied as a dynamical system. Other deviations affect much more, and in a way more difficult to apprehend, to the evolutionary dynamics and its equilibria, such as considering finite size populations, alternative learning/reproduction dynamics, or the non-universality of interactions among individuals. In this section we will choose this last point as our specific example, and analyze the consequences of relaxing the hypothesis that every player plays every other one. This hypothesis is needed to substitute the payoff earned by a player by what she would have obtained facing the average player of the population (an approach that has been traditionally used in physics under the name of mean-field approximation). However, interactions may not be universal after all, either because of spatial or temporal limitations. We will address both in what follows. The reader is referred to [15] for discussions of other perturbations.

### Spatial Perturbations

One of the reasons why maybe not all individuals interact with all others is that they could not possibly meet. In biological terms this may occur because the population is very sparsely distributed and every individual meets only a few others within its living range, or else in a very numerous population where it is impossible in practice to meet all individuals. In social terms, an alternative view is the existence of a social network or network of contacts that prescribes who interacts with whom.

This idea was first introduced in a famous paper by [16] on the evolutionary dynamics of the Prisoner's Dilemma on a square lattice. In the Prisoner's Dilemma two players simultaneously decide cooperate or to defect. Cooperation results in a benefit  $b$  to the recipient but incurs costs  $c$  to the donor ( $b > c > 0$ ). Thus, mutual cooperation pays a net benefit of  $R = b - c$  whereas mutual

defection results in  $P = 0$ . However, unilateral defection yields the highest payoff  $T = b$  and the cooperator has to bear the costs  $S = -c$ . It immediately follows that it is best to defect regardless of the opponents decision. For this reason defection is the evolutionarily stable strategy even though all individuals would be better off if all would cooperate (mutual cooperation is better than mutual defection because  $R > P$ ). Mutual defection is also the only Nash equilibrium of the game. All this translates into the following payoff matrix:

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \left( \begin{array}{cc} R & S \\ T & P \end{array} \right) \end{array} . \quad (9)$$

For this matrix to correspond to a Prisoner's Dilemma game, the ordering of payoffs must be  $T > R > P > S$ . As we will see below, other orderings define different games.

What Nowak and May did was to set the individuals on the nodes of a square lattice, where they played the game only with their nearest and next-nearest neighbors (Moore neighborhood). They ran simulations with the following dynamics: every individual played the game with her neighbors and collected the corresponding payoff, and afterwards she updated her strategy by imitating that of her most successful (in terms of payoff) neighbor. In their simulations, Nowak and May found that if they started with a population with a majority of cooperators, a large fraction of them remained cooperators instead of changing their behavior towards the ESS, namely, defection. The reason is that the structured interaction allowed cooperators to do well and avoid exploitation by defectors by grouping into clusters, inside which they interacted mostly with other cooperators, whereas defectors at the boundaries of those clusters, interacting mostly with other defectors, did not fare as well and therefore did not induce cooperators to defect.

This result is partly due to the imitation dynamics, which, if postulated to rule the evolution of a population of individuals that interact with all others, does not lead to the replicator equation. As we mentioned in the preceding section, the update rule for the strategies is arbitrary and can be chosen at will (preferably with some specific modelling in my mind). To reproduce the behavior of the replicator equation, a probabilistic rule has to be used [4], and with this rule the equilibrium is not changed and the population evolves to full defection. However, [16] opened the door to a number of more detailed studies that considered also different dynamical rules including the one corresponding to the replicator dynamics in which it was shown that the structure of a population definitely had a strong influence on the game equilibria.

One such study, perhaps the most systematic to date, was carried out by [5], who compared the equilibrium frequencies of cooperators and defectors in populations with and without spatial structuring (square lattices), finding two important results: First, including spatial extension has indeed significant effects on the equilibrium frequencies of cooperators and defectors. In some parameter regions spatial extension promotes cooperative behavior while inhibiting it in others; and, second, differences in the initial frequencies of cooperators are readily leveled out and hardly affect the equilibrium frequencies except for  $T < 1$ ,  $S < 0$ . This choice is not the Prisoner's Dilemma anymore, it corresponds to the so-called Stag Hunt game [22], and in the replicator dynamics is a bistable system where the initial frequencies determine the long term behavior, a feature that is generally preserved for the spatial setting. Of course, [5] also observed that the size of the neighborhood obviously affects the spreading speed of successful strategies. Interestingly, although the message seems to be that the strategy of cooperation is favored over the strategy of defection by the presence of a spatial structure, this is not always the case, and in games where the equilibrium population has a certain percentage of both types of strategists, the network of interactions makes the frequency of cooperators decrease [6]. Therefore, the effect of this perturbation is not all trivial and needs careful consideration.

All the results discussed so far correspond to a square lattice as substrate to define the interaction pattern, but this is certainly a highly idealized setup that can hardly correspond to any real, natural system. Recent studies have shown that the results also depend on the type of graph or network used. Thus, [21] have shown that in more realistic, heterogeneous populations, modeled by random graphs of different types, the sustainability of cooperation (implying the departure of the equilibrium predicted by the replicator equation) is simpler to achieve than in homogeneous populations, a result which is valid irrespective of the dilemma or game adopted as a metaphor of cooperation. Therefore, heterogeneity constitutes a powerful mechanism for the emergence of cooperation (and consequently an important perturbation of the dynamics), since even for mildly heterogeneous populations it leads to sizeable effects in the evolution of cooperation. The overall enhancement of cooperation obtained on single-scale and scale-free graphs [1] may be understood as resulting from the interplay of two mechanisms: The existence of many long-range connections in random and small-world networks [25], which precludes the formation of compact clusters of cooperators, and the heterogeneity exhibited by these networks, which opens a new

route for cooperation to emerge and contributes to enhance cooperation (which increases with heterogeneity), counteracting the previous effect. This result depends also on the intricate ties between individuals, even for the same class of graphs, features absent in the replicator dynamics.

We have thus seen that removing the hypothesis of universal interaction is a strong perturbation to equilibrium as understood from the replicator dynamics. There have been many other studies following those we can possibly review here; a recent, very comprehensive summary can be found in [23]. It must be realized that, intermixed with the effect of the network of interactions, the different dynamical rules one can think of have also a relevant influence on the equilibria. While we have not considered them as perturbations of the replicator dynamics here, because they do not behave as described by that equation even in the presence of universal interactions, it must be kept in mind that they do affect the equilibria and therefore they must be properly specified in any serious study of evolutionary game theory.

### Time Scales

Let us now come back to the situation in which there is no spatial structure, and every agent can in principle play the game against every other one. Afterwards, reproduction proceeds according to the payoff earned during the game stage. As we have already said, for large populations, this amounts to saying that every player gains the payoff of the game averaged in the current distribution of strategies. In terms of time scales, such an evolution corresponds to a regime in which reproduction-selection events take place at a much slower rate than the interaction between agents. However, these two time scales need not be different in general and, in fact, for many specific applications they can arguably be of the same order [7].

To study different rates of selection we can consider the following new dynamics [19,20,24]. There is a population with  $N$  players. A pair of individuals is randomly selected for playing, earning each one an amount of fitness according to the rules of the game. This game act is repeated  $s$  times, choosing a new random pair of players in each occasion.

Afterwards, selection takes place. Following [17], we have chosen Moran dynamics [13] as the most suitable to model selection in a finite population. This is necessary because the replicator equation is posed for continuous values of the populations and here we need to consider discrete values, i. e., individual by individual, in order to pinpoint the existing time scales. However, it can be shown

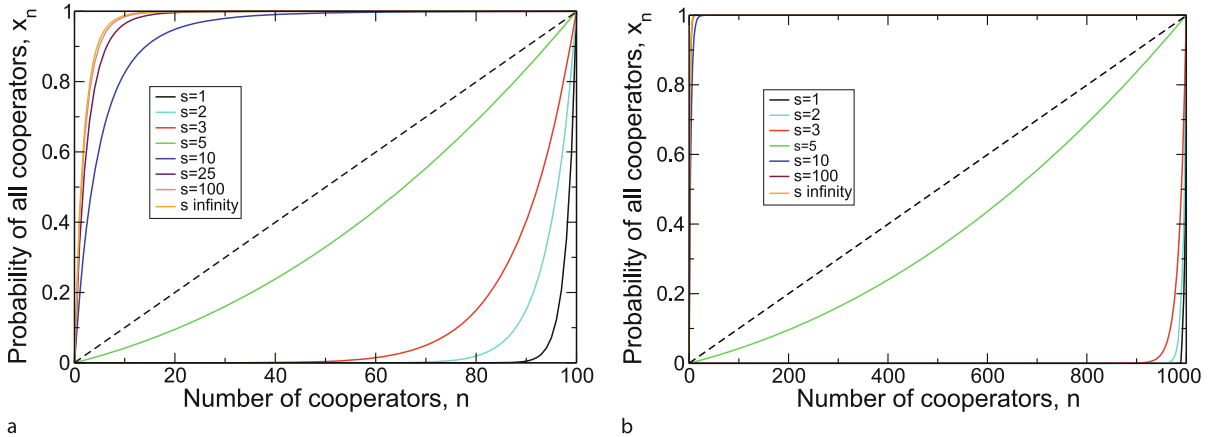
[19] that the equilibria of Moran dynamics are the same as those of the replicator equation, and in fact, the whole evolution is the same except for a rescaling of time. Moran dynamics is defined as follows: One individual among the population of  $N$  players is chosen for reproduction proportionally to its fitness, and its offspring replaces a randomly chosen individual. As the fitness of all players is set to zero before the following round of  $s$  games, the overall result is that all players have been replaced by one descendant, but the player selected for reproduction has had a reproductive advantage of doubling its offspring at the expense of the randomly selected player. It is worth noting that the population size  $N$  is therefore constant along the evolution.

The parameter  $s$  controls the time scales of the model, i. e. reflects the relation between the rate of selection and the rate of interaction. For  $s \ll N$  selection is very fast and very few individuals interact between reproduction events. Higher values of  $s$  represent proportionally slower rates of selection. Thus, when  $s \gg N$  selection is very slow and population is effectively well-mixed and we recover the behavior predicted by the replicator equation.

The most striking example of the influence of the selection is the so-called Harmony game, a trivial one that has henceforth never been studied, and that is determined by  $R > T > S > P$ . The *only* Nash equilibrium or ESS of this game is mutual cooperation, as it is obvious from the payoffs: The best option for both players is to cooperate, which yields the maximum payoff for each one. Let us denote by  $0 \leq n \leq N$  the number of cooperators present in the population, and look at the probability  $x_n$  of ending up in state  $n = N$  (i. e., all players cooperate) when starting in state  $n < N$ . For  $s = 1$  and  $s \rightarrow \infty$ , an exact, analytical expression for  $x_n$  can be obtained [19]. For arbitrary values of  $s$ , such a closed form cannot be found; however, it is possible to carry out a combinatorial analysis of the possible combinations of rounds and evaluate, numerically but exactly,  $x_n$ .

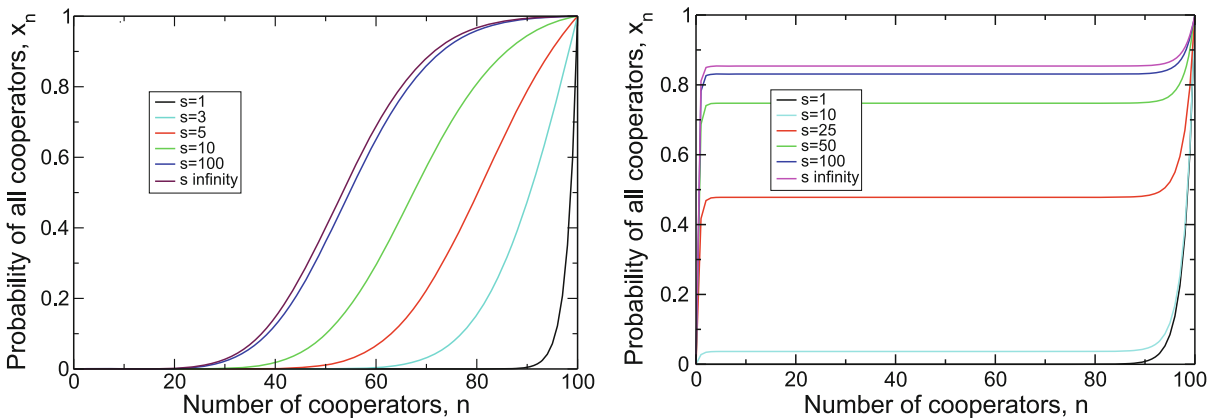
In Fig. 2a, we show that the rationally expected outcome of a population consisting entirely of cooperators is not achieved for small and moderate values of  $s$ , our selection rate parameter. For the smallest values, only when starting from a population largely formed by cooperators there is some chance to reach full cooperation; most of the times, defectors will eventually prevail and invade the whole population. This counterintuitive result may arise even for values of  $s$  comparable to the population size, by choosing suitable payoffs. Interestingly, the main result that defection is selected for small values of  $s$  does not depend on the population size  $N$ ; only details such as the shape of the curves (cf. Fig. 2b) are modified by  $N$ .





### Perturbation of Equilibria in the Mathematical Theory of Evolution, Figure 2

Probability of ending up with all cooperators starting from  $n$  cooperators,  $x_n$  for different values of  $s$ . **a** For the smallest values of  $s$ , full cooperation is only reached if almost all agents are initially cooperators. Values of  $s$  of the order of 10 show a behavior much more favorable to cooperators. In this plot, the population size is  $N = 100$ . **b** Taking a population of  $N = 1000$ , we observe that the range of values of  $s$  for which defectors are selected does not depend on the population size, only the shape of the curves does. Parameter choices are: Number of games between reproduction events,  $s$ , as indicated in the plots; payoffs for the Harmony game,  $R = 11$ ,  $S = 2$ ,  $T = 10$ ,  $P = 1$ . The dashed line corresponds to a probability to reach full cooperation equal to the initial fraction of cooperators and is shown for reference



### Perturbation of Equilibria in the Mathematical Theory of Evolution, Figure 3

**Left:** Same as Fig. 2 for the Stag-Hunt game. The probability of ending up with all cooperators starting from  $n$  cooperators,  $x_n$ , is very low when  $s$  is small, and as  $s$  increases it tends to a quasi-symmetric distribution around  $1/2$ . Payoffs for the Stag-Hunt game,  $R = 6$ ,  $S = 1$ ,  $T = 5$ ,  $P = 2$ . **Right:** Same as Fig. 2 for the Snowdrift game. The probability of ending up with all cooperators starting from  $n$  cooperators is almost independent of  $n$  except for very small or very large values. Small  $s$  values lead once again to selection of defectors, whereas cooperators prevail more often as  $s$  increases. Payoffs for the Snowdrift game,  $R = 1$ ,  $S = 0.35$ ,  $T = 1.65$ ,  $P = 0$ . Other parameter choices are: Population,  $N = 100$ ; number of games between reproduction events,  $s$ , as indicated in the plot

In the preceding paragraph we have chosen the Harmony game to discuss the effect of the rate of selection, but this effect is very general and appears in many other games. Consider the example of the already mentioned Stag-Hunt game [22], with payoffs  $R > T > P > S$ . This is the paradigmatic situation of game with two Nash equilibria in pure strategies, mutual cooperation and mutual defection, each one with its own basis of attraction in the

replicator equation framework (in general, which of these equilibria is selected has been the subject of a long argument in the past, and rationales for both of them can be provided [22]). As Fig. 3 (left) shows, simulation results for finite  $s$  are largely different from the curve obtained for  $s \rightarrow \infty$ : Indeed, we see that for  $s = 1$ , all agents become defectors except for initial densities close to 1. Even for values of  $s$  as large as  $N$  evolution will more

likely lead to a population entirely consisting of defectors.

Yet another example of the importance of the selection rate is provided by the Snowdrift game, with payoffs  $T > R > S > P$ . Figure 3 (right) shows that for small values of  $s$  defectors are selected for almost any initial fraction of cooperators. When  $s$  increases, we observe an intermediate regime where both full cooperation and full defection have nonzero probability, which, interestingly, is almost independent of the initial population. And, for large enough  $s$ , full cooperation is almost always achieved.

With these examples, it is clear that considering independent interaction and selection time scales may lead to highly non-trivial, counter-intuitive results. [19] showed that out of the 12 possible different games with two players, six were severely changed by the introduction of the game time scale, whereas the other six remained with the same equilibrium structure. Of course, the extent of the modifications of the replicator dynamics picture depends on the structure of the unperturbed phase space. Thus, rapid selection perturbations show up in changes of the asymptotically selected equilibria, i. e., of the asymptotically stable one, to changes of the basins of attraction of equilibria, or to suppression of long-lived metastable equilibria. As in the case of spatial perturbations, we are thus faced with a most relevant influence on the equilibria of the evolutionary game.

### Future Directions

As we have seen in this necessarily short excursion, the simplest mathematical models of evolution allow for a detailed, analytical study of their equilibria (which are supposed to represent stable states of populations) but, when leaving aside some of the hypothesis involved in the derivation of those simple models, the structure of equilibria may be seriously modified and highly counter-intuitive results may arise. We have not attempted to cover all possible perturbations but we believe we have provided evidence enough that their effect is certainly very relevant. When trying to bridge the gap between simple models and reality, other hypotheses will be broken, maybe more than one simultaneously, and subsequently the equilibria will be severely affected.

In the future, we believe that this line of research will undergo very interesting developments, particularly in the framework of evolutionary game theory, as the fitness landscape picture seems to be rather well understood and, on the other hand, is felt to be a much too simple model. In the case of evolutionary games, while some of the results we have collected here are analytical, there are many

others which are only numerical, in particular when perturbations depending on space (evolutionary game theory on graphs) are considered. A lot of research needs to be devoted in the next few years to understand this problem analytically, more so because there is now a “zoo” of results that are even hard to classify or interpret within a common basis. This will probably require a combined effort from different mathematical disciplines, ranging from discrete mathematics to dynamical systems through, of course, graph theory.

On the other hand, our examples have consisted of two-player, two-strategy, symmetric games, i. e., the simplest possible scenario. There are practically no results about games with more than two strategies or, even worse, with more than two players. In fact, even the classification of the phase portraits within the replicator equations for those situations is far from understood, the more so the higher the dimensionality of the problem. Much remains to be done in this direction. Asymmetrical games are a different story; for those, the replicator equation is not (6) anymore but rather one has to take into account that payoffs when playing as player 1 or player 2 are not the same, and the corresponding equation is more complicated. Again, this line of research is still in its infancy and awaiting for dedicated work.

Finally, a very interesting direction is the application of the results to problems in social or biological contexts. The evolutionary game theory community has been relying strongly on the predictions from the replicator equation which we now see may not agree with reality or at least with what occurs when some of its hypotheses are not fulfilled. This has led to a number of conundrums, particularly prominent among those being the problem of the emergence of cooperation. A recent study [24] have shown that, in a scenario described by the so-called Ultimatum game, taking into consideration the possible separation of time scales leads to results compatible with the experimental observations on human subjects, observations that the replicator equation is not able to reproduce. We envisage that similar results will be ubiquitous when trying to match the predictions of the replicator equation with actual systems or problems. Understanding the effect of perturbations in a comprehensive manner will then be the key to the fruitful development of the theory as a “natural” or “physical” one.

### Bibliography

#### Primary Literature

1. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512

2. Drossel B (2001) Biological evolution and statistical physics. *Adv Phys* 50:209–295
3. Fisher RA (1958) *The Genetical Theory of Natural Selection*, 2nd edn. Dover, New York
4. Gintis H (2000) *Game Theory Evolving*. Princeton University, Princeton
5. Hauert C (2002) Effects of space in  $2 \times 2$  games. *Int J Bifurc Chaos* 12:1531–1548
6. Hauert C, Doebeli M (2004) Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* 428:643–646
7. Hendry AP, Kinnison MT (1999) The pace of modern life: Measuring rates of contemporary microevolution. *Evolution* 53:1637–1653
8. Hofbauer J, Sigmund K (1998) *Evolutionary Games and Population Dynamics*. Cambridge University, Cambridge
9. Hofbauer J, Sigmund K (2003) Evolutionary game dynamics. *Bull Am Math Soc* 40:479–519
10. Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University, Cambridge
11. Maynard-Smith J (1982) *Evolution and the Theory of Games*. Cambridge University, Cambridge
12. Maynard-Smith J, Price GR (1973) The logic of animal conflict. *Nature* 246:15–18
13. Moran PAP (1962) *The Statistical Processes of Evolutionary Theory*. Clarendon, Oxford
14. Nash JF (1950) Equilibrium points in  $n$ -person games. *Proc Natl Acad Sci USA* 36:48–49
15. Nowak MA (2006) *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard Univ Press, Harvard
16. Nowak MA, May R (1992) Evolutionary games and spatial chaos. *Nature* 359:826–829
17. Nowak MA, Sasaki A, Taylor C, Fudenberg D (2004) Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428:646–650
18. Page K, Nowak MA (2002) A unified evolutionary dynamics. *J Theor Biol* 219:93–98
19. Roca CP, Cuesta JA, Sánchez A (2006) Time scales in evolutionary dynamics. *Phys Rev Lett* 97: art. no. 158701
20. Roca CP, Cuesta JA, Sánchez A (2007) The importance of selection rate in the evolution of cooperation. *Eur Phys J Special Topics* 143:51–58
21. Santos FC, Pacheco JM, Lenaerts T (2006) Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc Natl Acad Sci USA* 103:3490–3494
22. Skyrms B (2003) *The Stag Hunt and the Evolution of Social Structure*. Cambridge University, Cambridge
23. Szabó G, Fáth G (2007) Evolutionary games on graphs. *Phys Rep* 446:97–216
24. Sánchez A, Cuesta JA (2005) Altruism may arise by individual selection. *J Theor Biol* 235:233–240
25. Watts DJ, Strogatz SH (1998) Collective dynamics of “Small World” Networks. *Nature* 393:440–442
26. Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc 6th Int Cong Genet* 1:356–366

## Books and Reviews

- Nowak M, Sigmund K (2004) Evolutionary dynamics of biological games. *Science* 303:793–799

- Taylor PD, Jonker L (1978) Evolutionarily stable strategies and game dynamics. *J Math Biosci* 40:145–156
- von Neumann J, Morgenstern O (1944) *Theory of Games and Economic Behavior*. Princeton University Press, Princeton

## Perturbation of Systems with Nilpotent Real Part

TODOR GRAMCHEV

Dipartimento di Matematica e Informatica,  
Università di Cagliari, Cagliari, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Complex and Real Jordan Canonical Forms](#)

[Nilpotent Perturbation and Formal Normal Forms of Vector Fields and Maps Near a Fixed Point](#)

[Loss of Gevrey Regularity in Siegel Domains in the Presence of Jordan Blocks](#)

[First-Order Singular Partial Differential Equations](#)

[Normal Forms for Real Commuting Vector Fields](#)

[with Linear Parts Admitting Nontrivial Jordan Blocks](#)

[Analytic Maps near a Fixed Point in the Presence of Jordan Blocks](#)

[Weakly Hyperbolic Systems and Nilpotent Perturbations](#)

[Bibliography](#)

### Glossary

**Perturbation** Typically, one starts with an “initial” system  $S_0$ , which is usually simple and/or well understood. We perturb the system by adding a (small) perturbation  $R$  so that the new object becomes  $S_0 + R$ . In our context the typical examples for  $S_0$  will be systems of linear ordinary differential equations with constant coefficients in  $\mathbb{R}^n$  or the associated linear vector fields.

**Nilpotent linear transformation** Let  $A: \mathbb{K}^n \mapsto \mathbb{K}^n$  be a linear map, where  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{K} = \mathbb{C}$ . We call  $A$  nilpotent if there exists a positive integer  $r$  such that the  $r$ th iteration  $A^r$  become the zero map, in short  $A^r = 0$ .

**Gevrey spaces** Let  $\Omega$  be an open domain in  $\mathbb{R}^n$  and let  $\sigma \geq 1$ . The Gevrey space  $G^\sigma(\Omega)$  stands for the set of all functions  $f \in C^\infty(\Omega)$  such that for every compact subset  $K \subset\subset \Omega$  one can find  $C = C_{K,f} > 0$  such that

$$\sup_{x \in K} |\partial_x^\alpha f(x)| \leq C^{|\alpha|+1} \alpha!^\sigma \quad (1)$$

for all  $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{Z}_+^n$ ,  $\alpha! = \alpha_1! \dots \alpha_n!$ ,  $|\alpha| := \alpha_1 + \dots + \alpha_n$ . If  $\sigma = 1$  we recapture the space of real analytic functions in  $\Omega$  while the scale  $G^\sigma(\Omega)$ ,  $\sigma > 1$ , serves as an intermediate space between the real analytic functions and the set of all  $C^\infty$  functions in  $\Omega$ . By the Stirling formula one may replace  $\alpha!^\sigma$  by  $|\alpha|!^\sigma$ ,  $|\alpha|^{\sigma|\alpha|}$  or  $\Gamma(\sigma|\alpha|)$ , where  $\Gamma(z)$  stands for the Euler Gamma function cf. the book of Rodino [41] for more details on the Gevrey spaces.

One associates also Gevrey index to formal power series, namely, given a (formal) power series

$$f(x) = \sum_{\alpha} f_{\alpha} x^{\alpha}$$

this is in the formal Gevrey space  $G_f^{\sigma}(\mathbb{K}^n)$  if there exist  $C > 0$  and  $R > 0$  such that

$$|f_{\alpha}| \leq C^{|\alpha|+1} |\alpha|!^{\sigma-1} \quad (2)$$

for all  $\alpha \in \mathbb{Z}_+^n$ .

In fact, one can find in the literature another definition of the formal Gevrey spaces  $G_f^{\tau}$  of index  $\tau$ , namely replacing  $\sigma - 1$  by  $\tau$  (see e. g. Ramis [40]).

### Definition of the Subject

The main goal of this article is to dwell upon the influence of the presence (explicit and/or hidden) of nontrivial real nilpotent perturbations appearing in problems in Dynamical Systems, Partial Differential Equations and Mathematical Physics. Under the term nilpotent perturbation we will mean, broadly speaking, a classical linear algebra type setting: we start with an object (vector field or map near a fixed point, first-order singular partial differential equations, system of evolution partial differential equations) whose “linear part”  $A$  is semisimple (diagonalizable) and we add a (small) nilpotent part  $N$ . The problems of interest might be summarized as follows: are the “relevant properties” (in suitable functional framework) of the initial “object” stable under the perturbation  $N$ . If not, to classify, if possible, the novel features of the perturbed systems.

Broadly speaking, the cases when the instabilities occur are rare, they form some kind of exceptional sets. However, they appear in important problems (both in mathematical and physical contests) when degeneracies (bifurcations) occur.

### Introduction

We will focus our attention on topics where the presence of nontrivial Jordan blocks in the linear parts changes the properties of the original systems (i. e., instabilities occurs unless additional restrictions are imposed):

- (i) Convergence/divergence issues for the normal form theory of vector fields and maps near a singular (fixed) point in the framework of spaces of analytic functions and Gevrey classes.
- (ii) (Non)solvability for singular partial differential equations near a singular point.
- (iii) Cauchy problems for hyperbolic systems of partial differential equations with multiple characteristics.

Some basic features of the normal form theory for vector fields near a point will be recalled with an emphasis on the difficulties appearing in the presence of nontrivial Jordan blocks in the classification and computational aspects of the normal forms. For more details and various aspects of perturbation theory in Dynamics we refer to other articles in the Perturbation Theory Section of this Encyclopedia: cf. Bambusi ► [Perturbation Theory for PDEs](#), Gaeta ► [Non-linear Dynamics, Symmetry and Perturbation Theory](#) in, Gallavotti ► [Perturbation Theory](#), Broer ► [Normal Forms in Perturbation Theory](#), Broer and Hansmann ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#), Teixeira ► [Perturbation Theory for Non-smooth Systems](#), Verhulst ► [Perturbation Analysis of Parametric Resonance](#), Walcher ► [Perturbative Expansions, Convergence of](#).

We start by outlining some motivating examples.

Consider the nilpotent planar linear system of ordinary differential equations

$$\dot{x} = \begin{pmatrix} 0 & \varepsilon \\ 0 & 0 \end{pmatrix} x, \quad x(0) = x^0 \in \mathbb{R}^2,$$

where  $\varepsilon \in \mathbb{R}$ . The explicit solution is given by  $x_1(t) = x_1^0 + x_2^0 \varepsilon t$ ,  $x_2(t) = x_2^0$  and clearly the equilibrium  $(0, 0)$  is not stable if  $\varepsilon \neq 0$ . On the other hand, if  $U(x_1)$  is a smooth real valued analytic function, satisfying  $U(0) = 0$ ,  $U(x_1) > 0$  for  $x_1 \neq 0$ , then it is well known that for  $\varepsilon > 0$  the Newton equation

$$\dot{x} = \begin{pmatrix} 0 & \varepsilon \\ 0 & 0 \end{pmatrix} x - \begin{pmatrix} 0 \\ U'(x_1) \end{pmatrix}$$

is stable at  $(0, 0)$ .

Another example, which enters in the framework considered here, is given by a conservative (i. e. Hamiltonian) dynamical system perturbed by a friction term.

Next, we illustrate the influence of the real nilpotent perturbations in the realm of the normal form theory and in the general theory of singular partial differential equations. Consider the linear PDE

$$(x_1 + \varepsilon x_2) \partial_{x_1} u + x_2 \partial_{x_2} u - \sqrt{2} x_3 \partial_{x_3} u - u = f(x),$$

where  $\varepsilon \in \mathbb{R}$ , and  $f$  stands for a convergent power series in a neighborhood of the origin of  $\mathbb{R}^3$ , at least quadratic at  $x = 0$ . Such equations appear in the so-called homological equations for the reduction to Poincaré–Dulac linear normal form of systems of analytic ODEs having an equilibrium at the origin. It turns out that in the semisimple case  $\varepsilon = 0$  we can solve the equation in the space of convergent power series while for  $\varepsilon \neq 0$  (i. e., when a nontrivial Jordan block appears) the equation is solvable only formally, namely, divergent solutions appear. One is led in a natural way to study the Gevrey index of divergent solutions.

### Complex and Real Jordan Canonical Forms

We start by revisiting the notion of complex and real canonical Jordan forms.

Recall that each linear map is decomposed uniquely into the sum of a semisimple map and a nilpotent one. We state a classical result in linear algebra

**Lemma 1** *Let  $A$  be an  $n \times n$  matrix with real or complex entries. Then it is uniquely decomposed as*

$$A = A_s + A_{\text{nil}}, \quad (3)$$

where  $A_s$  is semisimple (i. e., diagonalizable over  $\mathbb{C}$ ) and  $A_{\text{nil}}$  is nilpotent, i. e.,  $A_{\text{nil}}^r = 0_{n \times n}$  for some positive integer  $r$ . Here  $0_{n \times n}$  stands for the zero  $n \times n$  matrix.

Denote by  $\text{spec}(A) = \{\lambda_1, \dots, \lambda_n\} \subset \mathbb{C}$  the set of all eigenvalues of  $A$  counted with their multiplicity.

The complex Jordan canonical form (JCF) is defined by the following assertion cf. Gantmacher [21]:

**Theorem 2** *Let  $A$  be an  $n \times n$  matrix over  $\mathbb{K}$ ,  $\mathbb{K} = \mathbb{C}$  or  $\mathbb{K} = \mathbb{R}$ . Then there exist positive integers  $m, k_1, \dots, k_m$ ,  $m \geq n$ ,  $k_1 + \dots + k_m = n$  and a matrix  $S \in GL(n; \mathbb{C})$  such that*

$$S^{-1}AS = J_A := \begin{pmatrix} \lambda_1^A I_{k_1} + N_{k_1} & 0_{k_1 \times k_2} & \dots & 0_{k_1 \times k_p} \\ 0_{k_2 \times k_1} & \lambda_2^A I_{k_2} + N_{k_2} & \dots & 0_{k_2 \times k_3} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{k_m \times k_1} & 0_{k_m \times k_2} & \dots & \lambda_m^A I_{k_m} + N_{k_m} \end{pmatrix} \quad (4)$$

where  $\lambda_1^A, \dots, \lambda_m^A$  are the eigenvalues of  $A$ , which need not all be distinct, and  $N_r$ , when  $r \geq 2$ , stands for the square  $r \times r$  matrix

$$\begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix} \quad (5)$$

with the convention  $N_1 = 0$ . Moreover, if  $A_{\text{nil}} \neq 0$ , i. e.,  $k_j \geq 2$  for at least one  $j \in \{1, \dots, m\}$ , then for every  $\varepsilon \in \mathbb{C} \setminus 0$  the matrix  $S(\varepsilon) \in GL(n; \mathbb{C}) = \text{diag}\{S_1(\varepsilon), \dots, S_m(\varepsilon)\}$ ,  $S_j(\varepsilon) = 1$  if  $k_j = 1$ ,  $S_j(\varepsilon) = \text{diag}\{1, \dots, \varepsilon^{k_j-1}\}$ , provided  $k_j \geq 2$ ,  $j = 1, \dots, m$ , satisfies the identity

$$S^{-1}(\varepsilon)AS(\varepsilon) = J_A(\varepsilon) = \begin{pmatrix} \lambda_1^A I_{k_1} + \varepsilon N_{k_1} & 0_{k_1 \times k_2} & \dots & 0_{k_1 \times k_p} \\ 0_{k_2 \times k_1} & \lambda_2^A I_{k_2} + \varepsilon N_{k_2} & \dots & 0_{k_2 \times k_3} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{k_m \times k_1} & 0_{k_m \times k_2} & \dots & \lambda_m^A I_{k_m} + \varepsilon N_{k_m} \end{pmatrix}. \quad (6)$$

One may define in an obvious way another JCF (lower triangular) replacing  $J$  by its transposed  $J^T$ .

Note that  $\lambda_k^A$ ,  $k = 1, \dots, m$ , are not necessarily distinct.

If  $\mu$  is an eigenvalue of  $A$  with algebraic multiplicity  $d$ , i. e., it is a zero of multiplicity  $d$  of the characteristic polynomial

$$P_A(\lambda) = |A - \lambda I|,$$

one can have different Jordan block structures. For example, a  $3 \times 3$  matrix with a triple eigenvalue  $\mu$  can be reduced to one of the three JCF:

- the matrix  $\mu I_3$ , i. e.,  $\mu$  does not admit nontrivial Jordan blocks;
- $\begin{pmatrix} \mu & 1 & 0 \\ 0 & \mu & 1 \\ 0 & 0 & \mu \end{pmatrix}$ ;
- $\begin{pmatrix} \mu & 0 & 0 \\ 0 & \mu & 1 \\ 0 & 0 & \mu \end{pmatrix}$

For higher dimensions the description of all JCF becomes more involved, cf. [3,21].

**Remark 3** We can choose  $|\varepsilon|$  arbitrarily small but never zero if the nilpotent part is nonzero. The columns of the conjugating matrix  $S$  are formed by eigenvectors and generalized eigenvectors. The smallness of  $|\varepsilon|$  leads in a natural way to view the presence of nontrivial nilpotent parts as a perturbation.

Next, if  $A$  is a real matrix, using the real and the imaginary parts of the eigenvectors for the complex eigenvalues, one introduces the real JCF.

**Theorem 4** *Let  $A \in M_{n \times n}(\mathbb{R})$ . Then there exist nonnegative integers  $p, q$ ,  $1 \leq p + 2q \leq n$ ,  $p$  (if  $p \geq 1$ ) positive integers  $k_1, \dots, k_p$ ,  $q$  (if  $q \geq 1$ ) positive integers  $\ell_1, \dots, \ell_q$*



satisfying  $k_1 + \dots + k_p + 2(\ell_1 + \dots + \ell_q) = n$ , and  $S \in SL(n; \mathbb{R})$ , such that

$$S^{-1}AS = \begin{pmatrix} J_A^R & 0_{k \times 2\ell} \\ 0_{2\ell \times k} & J_A^C \end{pmatrix} \quad (7)$$

with

$$J_A^R = \begin{pmatrix} \lambda_1^A I_{k_1} + \varepsilon_1^A N_{k_1} & 0_{k_1 \times k_2} & \dots & 0_{k_1 \times k_p} \\ 0_{k_2 \times k_1} & \lambda_2^A I_{k_2} + \varepsilon_2^A N_{k_2} & \dots & 0_{k_2 \times k_3} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{k_p \times k_1} & 0_{k_p \times k_2} & \dots & \lambda_p^A I_{k_p} + \varepsilon_p^A N_{k_p} \end{pmatrix}, \quad (8)$$

for some  $\lambda_j \in \mathbb{R}$ ,  $j = 1, \dots, p$ , and

$$J_A^C = \begin{pmatrix} D_1^A & 0_{2\ell_1 \times 2\ell_2} & \dots & 0_{2\ell_1 \times 2\ell_q} \\ 0_{2\ell_2 \times 2\ell_1} & D_2^A & \dots & 0_{2\ell_2 \times 2\ell_q} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2\ell_q \times 2\ell_1} & 0_{2\ell_q \times 2\ell_2} & \dots & D_q^A \end{pmatrix}, \quad (9)$$

with  $D_\mu^A$  being  $2\ell_\mu \times 2\ell_\mu$  matrices, written as  $\ell_\mu \times \ell_\mu$  block matrices of  $2 \times 2$  matrices of the following  $2 \times 2$  block matrix form:

$$D_\mu^A = \begin{pmatrix} \alpha_\mu & -\beta_\mu \\ \beta_\mu & \alpha_\mu \end{pmatrix} I_{\ell_\mu}(2) + \begin{pmatrix} \gamma_\mu^A & -\delta_\mu^A \\ \delta_\mu^A & \gamma_\mu^A \end{pmatrix} N_\mu(2), \quad (10)$$

for some  $\alpha_\mu, \beta_\mu \in \mathbb{R}$ ,  $\beta_\mu \neq 0$ , with  $\alpha_k \pm \beta_k i \in \text{spec}(A)$ . Here,  $I_k(2) = \text{diag}\{I_2, \dots, I_2\}$  denotes the  $2k \times 2k$  matrix written as a  $k \times k$  matrix with  $2 \times 2$  block matrices while  $N_k(2)$  stands for the following  $2k \times 2k$  nilpotent matrix written as  $k \times k$  matrix with  $2 \times 2$  block matrices as entries:

$$N_k(2) = \begin{pmatrix} 0_{2 \times 2} & I_2 & 0_{2 \times 2} & \dots & 0_{2 \times 2} \\ 0_{2 \times 2} & 0_{2 \times 2} & I_2 & \dots & 0_{2 \times 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{2 \times 2} & 0_{2 \times 2} & 0_{2 \times 2} & \dots & I_2 \\ 0_{2 \times 2} & 0_{2 \times 2} & 0_{2 \times 2} & \dots & 0_{2 \times 2} \end{pmatrix} \quad (11)$$

and  $0_{r \times s}$  stands for the zero  $r \times s$  matrix.

The smallness of the parameter  $\varepsilon$  and the explicit form of the conjugating matrices  $S(\varepsilon)$  are instrumental in showing some useful estimates for the study of the dynamics of the linear maps  $A$  which are not semisimple. Let  $r(A) := \max\{|\lambda| : \lambda \in \text{spec}(A)\}$  (the spectral radius of  $A$ ). Then the following estimate, useful in different branches of Dynamical Systems, holds (cf. [27])

**Lemma 5** For every  $\eta > 0$  there exists a norm in  $\mathbb{R}^n$  such that  $\|A\| \leq r(A) + \eta$ . Moreover, if

$$\{\lambda \in \text{spec}(A) : |\lambda| = r(A) \text{ do not admit Jordan blocks}\}$$

one has  $\|A\| = r(A)$  for some norm. In particular, if  $A$  is semisimple, the last conclusion holds.

Next, consider the linear autonomous systems of ordinary differential equations

$$\dot{x} = Ax. \quad (12)$$

where  $A \in M_n(\mathbb{R})$ . We recall that if  $x(0) = \xi \in \mathbb{R}^n$ , then the unique solution is defined by

$$x(t) = \exp(tA)\xi := \sum_{k=0}^{\infty} \frac{t^k A^k}{k!} \xi, \quad (13)$$

e.g., cf. Arnold [3], Coddington and Levinson [14].

We exhibit an assertion where the structure of the real nilpotent perturbation  $A_{\text{nil}}$  in the linear part plays a crucial role for the stability for  $t \geq 0$  of the solutions of the linear system (12). We recall that the origin is stable if for every  $\varepsilon > 0$  one can find  $\delta > 0$  such that  $\|\xi\| < \delta$  implies  $\|\exp(tA)\xi\| < \varepsilon$  for  $t \geq 0$ .

**Proposition 6** The zero solution of (12) is stable for  $t \geq 0$  if and only if the following two conditions hold:

- i)  $\text{spec}(A) \subset \{\lambda \in \mathbb{C} : \text{Re } \lambda \leq 0\}$ ;
- ii) if  $\lambda \in \text{spec}(A)$  and  $\text{Re } \lambda = 0$  then  $\lambda$  does not admit a nontrivial Jordan block.

On the other hand, the origin is asymptotically stable for  $t \rightarrow +\infty$  if and only if

$$\text{spec}(A) \subset \{\lambda \in \mathbb{C} : \text{Re } \lambda < 0\}.$$

The proof is straightforward in view of the explicit formula for the exponent of the matrix  $\exp(tA)$  by means of the Jordan canonical form.

In particular, we get

**Corollary 7** Let  $A$  be a real matrix such that all eigenvalues lie on the pure imaginary axis. Then the zero solution of (12) is stable if and only if  $A$  is semisimple (i.e.,  $A_{\text{nil}} = 0$ ).

### Nilpotent Perturbation and Formal Normal Forms of Vector Fields and Maps Near a Fixed Point

Normal form theory (originating back to Poincaré's thesis) has proven to be one of the most useful tools for the local

analysis of dynamical systems near an equilibrium (singular) point  $x^0$  for autonomous systems of ODE

$$\dot{x} = X(x) \quad (14)$$

or the associated vector field

$$\widetilde{X}(x) = \langle X(x), \partial_x \rangle = \sum_{j=1}^n X_j(x) \partial_{x_j} \quad (15)$$

(see [3,7,9,12,19,20,43], and the references therein). Without loss of generality (after a translation) one may assume that  $x^0$  coincides with the origin and write

$$\begin{aligned} X(x) &= Ax + R(x), \quad A = \nabla X(0), \\ R(x) &= O(|x|^2), \quad |x| \rightarrow 0. \end{aligned} \quad (16)$$

Denote by  $\text{spec}(A) = \{\lambda_1, \dots, \lambda_n\}$  the spectrum of  $A$ . The basic idea, going back to Poincaré, is to find a (formal) change of the coordinates defined as a (at least) quadratic perturbation of the identity

$$x = u(y) = y + v(y), \quad (17)$$

which transforms  $\widetilde{X}$  into a new vector field  $\widetilde{Y}(y)$  which has a “simpler” form (Poincaré–Dulac normal form).

The original idea of Poincaré regards the possibility to linearize  $\widetilde{X}$ , i. e.,  $\widetilde{Y}(y) = Ay$ . Straightforward calculations show that the linearization of  $\widetilde{X}$  means that  $v(y)$  satisfies (at least formally) a system of first order semilinear partial differential equations, called the system of the homological (difference) equations

$$L_A v(y) = R(y + v(y)) \quad (18)$$

In fact, it is the system above where the first substantial technical difficulty appears if the nilpotent part  $A_{\text{nil}}$  is not zero, i. e., the matrix  $A$  is not diagonalizable. Indeed, if  $A$  is diagonalizable and we choose (after a linear change of the variables in  $\mathbb{C}^n$ )  $A = \text{diag}\{\lambda_1, \dots, \lambda_n\}$ , then the system (18) is written as

$$\sum_{k=1}^n \lambda_k \partial_{y_k} v_j(y) - \lambda_j v_j = R_j(y + v(y)), \quad j = 1, \dots, n \quad (19)$$

We recall that  $\widetilde{X}$  (or  $\text{spec}A$ ) is said to be in the Poincaré domain (respectively, Siegel domain) if the convex hull of  $\{\lambda_1, \dots, \lambda_n\}$  in the complex plane does not contain (respectively, contains) 0. Further,  $\text{spec}(A)$  is called nonresonant iff

$$\langle \lambda, \alpha \rangle - \lambda_j \neq 0, \quad j = 1, \dots, n, \quad \alpha \in \mathbb{Z}_+^n(2), \quad (20)$$

where  $\langle \lambda, \alpha \rangle = \sum_{j=1}^n \lambda_j \alpha_j$ ,  $\mathbb{Z}_+^n(2) := \{\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{Z}_+^n : |\alpha| := \alpha_1 + \dots + \alpha_n \geq 2\}$ .

By the Poincaré–Dulac theorem, if  $\text{spec}(A)$  is in the Poincaré domain, then there are at most finitely many resonances and there exists a convergent transformation (in some neighborhood of the singular point) which reduces  $\widetilde{X}$  to a (finitely resonant) normal form.

**Theorem 8** *Let the linear part  $A$  of the complex (respectively, real) field above be nonresonant. Then the vector field is formally linearizable by a complex (respectively, real) transformation.*

Denote by

$$\text{Res}_j^A = \{\alpha \in \mathbb{Z}_+^n(2) : \lambda_j = \langle \lambda, \alpha \rangle\}, \quad j = 1, \dots, n.$$

Clearly the nonresonance hypothesis is equivalent to  $\text{Res}_j^A = \emptyset$  for  $j = 1, \dots, n$ .

Additional technical complications appear if we consider real vector fields.

**Theorem 9** *Every formal vector field with a singular point at the origin is transformed by a formal complex change of variables to a field of the form*

$$\langle J_A z, \partial_z \rangle + \sum_{j=1}^n \sum_{\alpha \in \text{Res}_j^A} q_{j,\alpha} z^\alpha \partial_{z_j} \quad (21)$$

The coefficients  $q_{j,\alpha}$  may be complex even though the original vector field is real.

We note that if the linear part is nilpotent, i. e.,  $\text{spec}(A) = \{0\}$ , then the theorem above gives no simplification. In that case Belitskii [7] has classified completely the formal normal forms.

Let  $A$  be a nilpotent matrix (i. e. the semisimple part  $A_s$  is the zero matrix). The Poincaré–Dulac NF does not provide any information. The following theorem is due to Belitskii [7] (see also Arnold and Ilyashenko [4]).

**Theorem 10** *Let  $A$  be a nilpotent matrix and let  $X(x)$  be a formal vector field with a linear part given by  $Ax$ . Then  $X$  is transformed by a formal complex change of variables  $x \mapsto z$  to a field of the form*

$$\langle J_A z, \partial_z \rangle + \langle B(z), \partial_z \rangle \quad (22)$$

with  $B(z)$  being at least quadratic near the origin, where the nonlinear vector field  $\langle B(z), \partial_z \rangle$  commutes with  $\langle J_A^* z, \partial_z \rangle$ . (Here  $*$  stands for the Hermitian conjugation).

We point out that another important problem is the computation of the normal form. Here the presence of the Jordan blocks leads to substantial difficulties.

The description and the computation of nilpotent normal forms, based on an algebraic approach and the systematic use of the theory of invariance, with particular emphasis on the equivalence between different normal forms, has been developed in a body of papers (see [16,36,37], and the references therein).

Finally, we mention that nilpotent perturbations appear in the classification and Casimir invariants of Lie–Poisson brackets that are formed by Lie algebra extensions for physical systems admitting Hamiltonian structure to such brackets (e. g. cf. Thiffeault and Morison [49]).

### Loss of Gevrey Regularity in Siegel Domains in the Presence of Jordan Blocks

The convergence question in the Siegel domain is more difficult since small divisors appear. In a fundamental paper Bruno [9] succeeded in proving a deep result of the following type: a formal normal form is convergent under an (optimal) arithmetic condition on the small divisors  $|\langle \lambda, \alpha \rangle - \lambda_j|^{-1}$  and a condition on the formal normal form, called the  $A$  condition. It should be pointed out that while in the original paper [9] the condition  $A$  allows in some cases nontrivial Jordan blocks of the linear part, in the subsequent works the linear part  $A$  is required to be semisimple (diagonalizable).

We recall that for the finitely smooth and  $C^\infty$  local normal forms the presence of the real nilpotent part does not influence the convergence: e. g., in the Sternberg theorem (cf. [45]) the small divisors and the presence of Jordan blocks play no role (cf. [7], where many other references can be found).

Little is known about convergence–divergence problems in the analytic category if  $\text{spec}(A)$  is in the Siegel domain and  $A$  is *not* semisimple.

Even in the cases where the presence of the nilpotent part does not influence the assertions, the proofs become more involved. Here we outline various aspects of normal forms when nilpotent perturbations are present. We stress especially the real case.

The combined influence of the Jordan blocks and the small divisors on the convergence of the formal linearizing transformation for analytic vector fields in  $\mathbb{R}^3$  has been studied in Gramchev [23] (see also [53] for some examples). Let  $n = 3$  and consider a nonsemisimple linear part  $A \in GL(3; \mathbb{K})$  satisfying  $\text{spec}(A) = \{\lambda, \mu, \mu\}$ ,  $\lambda \neq \mu$ . This means that we can reduce  $A$  to the  $\varepsilon$  Jordan normal form

$$A_\varepsilon = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \mu & \varepsilon \\ 0 & 0 & \mu \end{pmatrix}, \quad \varepsilon \neq 0. \quad (23)$$

We recall that one can make  $|\varepsilon|$  arbitrarily small by linear change of the variables (but never 0). Then  $\text{spec}(A)$  is nonresonant and in the Siegel domain iff

$$\rho := \frac{\lambda}{\mu} < 0, \quad \rho \notin \mathbb{Q}. \quad (24)$$

The typical example is a real vector field with a linear part given by

$$A_\varepsilon^0 = \begin{pmatrix} \rho & 0 & 0 \\ 0 & 1 & \varepsilon \\ 0 & 0 & 1 \end{pmatrix}, \quad \varepsilon \neq 0. \quad (25)$$

One observes that such real vector fields are nonresonant and hyperbolic and therefore, by the Chen theorem, linearizable by smooth transformations.

We recall that an irrational number  $\rho$  is said to be diophantine of order  $\tau > 0$ , and write  $\rho \in D(\tau)$ , if there exist  $C > 0$  such that

$$\min_{p \in \mathbb{Z}} |q\rho + p| \geq \frac{C}{q^\tau}, \quad q \in \mathbb{N}. \quad (26)$$

By a classical result in number theory  $D(\tau) \neq \emptyset$  iff  $\tau \geq 1$ . An irrational number  $\rho$  is called Liouville iff it is not diophantine.

Given an irrational number  $\rho$  we set

$$\tau_0 = \tau_0(\rho) = \inf\{\tau > 0: \text{ such that } \rho \in D(\tau)\}, \quad \text{before} \quad (27)$$

with the convention  $\tau_0 = +\infty$  if  $\rho$  is a Liouville number.

**Theorem 11** *Let  $\text{spec}(A)$  be in the Siegel domain. Then  $L_A$  is not solvable in the space of convergent power series, namely, we can find RHS  $f$  which is analytic but the unique formal power series solution is divergent. Moreover, we can always find a convergent RHS  $f$  such that the unique formal solution  $u$  satisfies*

$$u \notin \bigcup_{1 \leq \sigma < 2 + \tau_0} G_f^\sigma(\mathbb{K}^3) \quad (28)$$

*In particular, if  $\tau_0 = +\infty$ , then*

$$u \notin G^\sigma, \quad \sigma \geq 1. \quad (29)$$

In case a diophantine condition is satisfied, estimates on the Gevrey character of the divergent power series are derived.

The nonsolvability of the LHE in the spaces of the convergent power series does not exclude a priori that the vector field is linearizable by analytic transformations.

However, using a fundamental result of Pérez Marco [39] one can state the following assertion for hyperbolic vector fields which are not in the Poincaré domain.

**Theorem 12** *Let  $A$  be a real  $n \times n$  matrix which is hyperbolic, not semisimple and is not in the Poincaré domain. Let  $R(x) = (R_1(x), \dots, R(x))$  be a real valued analytic function near the origin, at least quadratic near  $x = 0$ , i.e.  $R(0) = 0$ ,  $\nabla R(0) = 0$ . Then there exists  $\eta_0 > 0$  such that for almost all  $\eta \in ]0, \eta_0]$  in the sense of the Lebesgue measure, the vector field defined by*

$$X_\varepsilon(x) = Ax + \eta R(x) \quad (30)$$

is not linearizable by convergent transformations.

**Remark 13** In fact, the result can be made more precise, using the notion of capacity instead of the Lebesgue measure and allowing polynomial dependence on  $\eta$  (cf. [39]).

Another direction where the presence of real nilpotent perturbations in the linear parts presents challenging obstacles is the study of the dynamics in a neighborhood of a fixed point carried out via a normalization up to finite order and the issue of optimal truncation

$$\sum_{|\alpha| \leq N_{\text{opt}}} u_\alpha x^\alpha$$

of normal form transformations for analytic vector fields. In an impressive work Iooss and Lombardi [33] demonstrate in particular that for large classes of real analytic vector fields with semisimple linear parts one can truncate the formal normal form transformation for  $|x| \leq \delta$ ,  $\delta_0 > 0$  arbitrarily small, in such a way that the remainder in the normal form  $R(x)$  satisfies the following estimates

$$\sup_{|x| \leq \delta} |R_{N_{\text{opt}}}(x)| \leq M \delta^2 \exp\left(-\frac{w}{\delta^b}\right) \quad (31)$$

where  $b = 1 + \tau$ . Here either  $\tau > n - 1$ , in which case  $\tau \neq 0$  is the diophantine index of the small divisor type estimates modulo the resonance set, namely for some  $\gamma > 0$  the following estimates hold for the eigenvalues  $\lambda_1, \dots, \lambda_n$  of  $A$

$$|\langle \lambda, \alpha \rangle - \lambda_j| \geq \frac{\gamma}{|\alpha|^\tau}, \quad \text{if } \langle \lambda, \alpha \rangle - \lambda_j \neq 0 \quad (32)$$

for  $j = 1, \dots, n$ ,  $\alpha \in \mathbb{Z}_+^n(2)$ , or  $\tau = 0$  and  $\lambda$  satisfies the nonresonant type estimates

$$|\langle \lambda, \alpha \rangle - \lambda_j| \geq \gamma, \quad \text{if } \langle \lambda, \alpha \rangle - \lambda_j \neq 0 \quad (33)$$

for  $j = 1, \dots, n$ ,  $\alpha \in \mathbb{Z}_+^n(2)$ .

The question of the validity of such results for analytic vector fields with nonsemisimple linearization is far more intricate. Iooss and Lombardi [33] give two examples of non-semisimple linearizations (nilpotent perturbations of size 2 and 3) for which the result is still true. The question remains totally open for other non-semisimple linearizations.

### First-Order Singular Partial Differential Equations

This section deals with the study of formal power series solutions to singular linear first-order partial differential equations with analytic coefficients of the form

$$\sum_{j=1}^d a_j(x) \partial_{x_j} u(x) + b(x) u(x) = f(x), \quad (34)$$

where  $a_j(x)$  (with  $j = 1, \dots, d$ ),  $b(x)$  and the right-hand side  $f(x)$  are analytic in a neighborhood of the origin of  $\mathbb{C}^d$ , and  $a_j(0) = 0$  for  $j = 1, \dots, d$ .

In an interesting paper Hibino [29] allows a Jacobian matrix  $\nabla a(0)$  without a Poincaré condition. More precisely, the Jacobi matrix at the origin can be reduced via a conjugation with a nonsingular matrix  $S$  to the following form: for some nonnegative integers,  $m, d, p$ , and  $d$  positive integers  $k_j \geq 2$  (with  $j = 1, \dots, d$ ), if  $d \geq 1$ , such that  $m + r_1 + \dots + r_d + p = n$ , one can write as follows:

$$S^{-1} \nabla a(0) S = \begin{pmatrix} A & & & \\ & N_{r_1} & & \\ & & \ddots & \\ & & & N_{r_d} \\ & & & & 0_{p \times p} \end{pmatrix} \quad (35)$$

where  $N_r$ ,  $r \geq 2$ , stands for the  $r \times r$  nilpotent Jordan block (5) and  $A$  is an  $m \times m$  satisfying the Poincaré condition (the convex hull in  $\mathbb{C}$  of the eigenvalues  $\lambda_1, \dots, \lambda_m$  of  $A$  does not contain the origin, provided  $m \geq 1$ ).

One observes that the hypothesis  $d \geq 1$  implies that  $\nabla a(0)$  does not satisfy the Poincaré condition.

The fundamental hypothesis on the zero order term  $b(x)$  reads as follows

$$\left| \sum_{r=1}^m \lambda_r \alpha_r + b(0) \right| \neq 0, \quad \alpha \in \mathbb{Z}_+^m \quad \text{if } m \geq 1 \quad (36)$$

$$|b(0)| \neq 0 \quad \text{if } m = 0 \quad (37)$$

In particular, by (37) one gets that necessarily  $b(0) \neq 0$ .

It should be stressed that the classes of singular partial differential equations above do not capture the systems of homological equations when small divisors occur, but

they outline some interesting features in the presence of nontrivial Jordan blocks even in the lack of small divisors phenomena.

Set  $\tau_0 = \max_{\ell=1,\dots,d} r_\ell$  if  $d \geq 1$ . Then the main result in [29] reads as follows

**Theorem 14** *Under the conditions (35)–(37) for every RHS*

$$f(x) = \sum_{\alpha \in \mathbb{Z}_+^n} f_\alpha x^\alpha$$

which converges in a neighborhood of the origin the equation (34) has a unique formal solution which belongs to the formal Gevrey space  $G_f^\sigma(\mathbb{K}^n)$  with

$$\sigma = \begin{cases} 2\tau_0 & \text{if } d \geq 1 \\ 2 & \text{if } d = 0, p \geq 1 \\ 1 & \text{if } d = p = 0 \end{cases} . \quad (38)$$

The Gevrey index is determined by a Newton polyhedron, a generalization of the notion of the Newton polygon for singular ordinary differential equations. The arguments of the proof rely on subtle Gevrey combinatorial estimates.

Extensions for first-order quasilinear singular partial differential equations are done by Hibino [31].

It should be pointed out that the loss of Gevrey regularity comes not only from the nilpotent Jordan blocks, but from the nonlinear (at least quadratic) terms in  $a(x)$  as well. Indeed, for the equation

$$(1 - x_1^2 \partial_{x_1} - x_2^2 \partial_{x_2})u(x) = x_1 + x_2$$

the unique formal solution is defined as follows

$$u(x) = \sum_{j=0}^{\infty} (j-1)!(x_1^j + x_2^j)$$

cf. [29]. For further investigations on the loss of Gevrey regularity for solutions of singular ordinary differential equations of irregular type see Gramchev and Yoshino [26]. For characterizations of the Borel summability of a divergent formal power series solution of classes of first-order linear singular partial differential equation of nilpotent type see [30] and the references therein.

Solvability in classical Sobolev spaces and Gevrey spaces for linear systems of singular partial differential equations with real coefficients in  $\mathbb{R}^n$  with nontrivial real Jordan blocks are derived by Gramchev and Tolis [24].

## Normal Forms for Real Commuting Vector Fields with Linear Parts Admitting Nontrivial Jordan Blocks

The main goal of this section is to exhibit the influence of nontrivial linear nilpotent parts for the simultaneous reduction to convergent normal forms of commuting vector fields with a common fixed point.

Consider a family of commuting  $n \times n$  matrices  $A_1, \dots, A_d$ . If all matrices are semisimple, then they can be simultaneously diagonalized over  $\mathbb{C}$  or put into a block-diagonal form over  $\mathbb{R}$ , if  $A_1, \dots, A_d$  are real, by a linear transformation  $S$  (e.g., [12,21]). This property plays a crucial role in the study of the normal forms of commuting vector fields with semisimple linear parts (cf. [46,47]).

However, if the matrices have nontrivial nilpotent parts, then it is not possible to transform simultaneously  $A_1, \dots, A_d$  in Jordan canonical forms. This is a consequence of the characterization of the centralizer of a matrix in a JCF (see [3,21]).

Apparently the first examples of simultaneous reduction to (formal) normal forms of commuting vector fields with nonsemisimple linear parts are due to Cicogna and Gaeta [12]) for two commuting vector fields using the set up of the symmetries. More precisely, first, the definition of Semisimple Joint Normal Form (SJNF) is introduced: let

$$X = \langle f(x), \partial_x \rangle, \quad Y = \langle g(x), \partial_x \rangle,$$

$A = \nabla f(0)$ ,  $B = \nabla g(0)$  and  $f(x) = Ax + F(x)$ ,  $g(x) = Bx + G(x)$ . Then  $X$  and  $Y$  are said to be in SJNF if both  $F$  and  $G$  belong to  $\text{Ker}(\mathcal{A}_s) \cap \text{Ker}(\mathcal{B}_s)$ . Here  $\mathcal{A}_s$  stands for the homological operator associated to the semisimple part  $A_s$  of  $A$ .

Next,  $X, Y$  are in  $X$ -Joint NF iff

$$F \in \text{Ker}(\mathcal{A}^+) \cap \text{Ker}(\mathcal{B}_s), \quad \text{and} \quad g \in \text{Ker}(\mathcal{A}_s) \cap \text{Ker}(\mathcal{B}_s).$$

Clearly in  $Y$ -JNF the role of  $f$  and  $g$  is reversed.

The main assertion is:

**Theorem 15** *Let  $[X, Y] = 0$ . Then  $X$  and  $Y$  can be reduced to a SJNF by means of a formal change of the variables. They can also be reduced (formally) to  $X$ -Joint or  $Y$ -Joint NF.*

**Example 16**

$$A = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & \varepsilon \\ 0 & 0 & \lambda \end{pmatrix} \quad B = \begin{pmatrix} p & 0 & q \\ r & s & t \\ 0 & 0 & s \end{pmatrix}, \quad (39)$$



where  $\lambda, \varepsilon, p, q, r, s, t \in \mathbb{C} \setminus 0$ . Then  $A$  and  $B$  commute but it is impossible, in general, to reduce  $B$  to the Jordan block structure of  $A$ , preserving that of  $A$ .

We point out that the problem of finding conditions guaranteeing simultaneous reduction of commuting matrices to Jordan normal forms is related to questions in the Lie group theory (e. g., see Chap. IV of [12] and the references therein).

Next, we outline recent results on simultaneous reductions to normal forms of commuting analytic vector fields admitting nontrivial Jordan blocks in their linear parts following Yoshino and Gramchev [54].

Let  $\mathbb{K}$  be  $\mathbb{K} = \mathbb{C}$  or  $\mathbb{K} = \mathbb{R}$ , and  $B = \infty$ ,  $B = \omega$  or  $B = k$  for some  $k > 0$ . Let  $\mathcal{G}_B^n$  denote a  $d$ -dimensional Lie algebra of germs at  $0 \in \mathbb{K}^n$  of  $C^B$  vector fields vanishing at 0. Let  $\rho$  be a germ of singular infinitesimal  $\mathbb{K}^d$ -actions of class  $C^B$  ( $d \geq 2$ )

$$\rho: \mathbb{K}^d \longrightarrow \mathcal{G}_B^n. \quad (40)$$

Denote by  $\text{Act}^B(\mathbb{K}^d: \mathbb{K}^n)$  the set of germs of singular infinitesimal  $\mathbb{K}^d$ -actions of class  $C^B$  at  $0 \in \mathbb{K}^n$ . By choosing a basis  $e_1, \dots, e_d \in \mathbb{K}^n$ , the infinitesimal action can be identified with a  $d$ -tuple of germs at 0 of commuting vector fields  $X^j = \rho(e_j)$ ,  $j = 1, \dots, d$  (cf. [18,46,47,55]). We can define, in view of the commutativity relation, the action

$$\begin{aligned} \tilde{\rho}: \mathbb{K}^d \times \mathbb{K}^n &\longrightarrow \mathbb{K}^n, \\ \tilde{\rho}(s; z) &= X_{s_1}^1 \circ \dots \circ X_{s_d}^d(z) = X_{s_{\sigma_1}}^{\sigma_1} \circ \dots \circ X_{s_{\sigma_d}}^{\sigma_d}(z), \quad (41) \\ s &= (s_1, \dots, s_d), \end{aligned}$$

for all permutations  $\sigma = (\sigma_1, \dots, \sigma_d)$  of  $\{1, \dots, d\}$ , where  $X_t^j$  denotes the flow of  $X^j$ . We denote by  $\rho_{\text{lin}}$  the linear action formed by the linear parts of the vector fields defining  $\rho$ .

A natural question is to investigate necessary and sufficient conditions for the linearization of  $\rho$  (allowing nilpotent perturbations in the linear parts) namely, whether there exists a  $C^B$  diffeomorphism  $g$  preserving 0 such that  $g$  conjugates  $\tilde{\rho}$  and  $\tilde{\rho}_{\text{lin}}$

$$\tilde{\rho}(s; g(z)) = g(\tilde{\rho}_{\text{lin}}(s, z)) \quad (s, z) \in \mathbb{K}^d \times \mathbb{K}^n. \quad (42)$$

It is well known (e. g., cf. [35]) that there exists a positive integer  $m \leq n$  such that  $\mathbb{K}^n$  is decomposed into a direct sum of  $m$  linear subspaces invariant under all  $A^\ell = \nabla X_\ell(0)$  ( $\ell = 1, \dots, d$ ):

$$\begin{aligned} \mathbb{K}^n &= \mathbb{I}^{s_1} + \dots + \mathbb{I}^{s_m}, \quad \dim \mathbb{I}^{s_j} = s_j, \quad j = 1, \dots, m, \\ s_1 + \dots + s_m &= n. \end{aligned} \quad (43)$$

The matrices  $A^1, \dots, A^d$  can be simultaneously brought into upper triangular form, and we write again  $A^\ell$  for the matrices

$$A^\ell = \begin{pmatrix} A_1^\ell & 0_{s_1 \times s_2} & \dots & 0_{s_1 \times s_m} \\ 0_{s_2 \times s_1} & A_2^\ell & \dots & 0_{s_2 \times s_m} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{s_m \times s_1} & 0_{s_m \times s_2} & \dots & A_m^\ell \end{pmatrix}, \quad \ell = 1, \dots, d. \quad (44)$$

If  $\mathbb{K} = \mathbb{C}$ , the matrix  $A_j^\ell$  is given by

$$A_j^\ell = \begin{pmatrix} \lambda_j^\ell & A_{j,12}^\ell & \dots & A_{j,1s_j}^\ell \\ 0 & \lambda_j^\ell & \dots & A_{j,2s_j}^\ell \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_j^\ell \end{pmatrix}, \quad (45)$$

with  $\lambda_j^\ell, A_{j,v\mu}^\ell \in \mathbb{C}$ ,  $\ell = 1, \dots, d$ ,  $j = 1, \dots, m$ . On the other hand, if  $\mathbb{K} = \mathbb{R}$ , then we have, for every  $1 \leq j \leq m$  two possibilities: firstly, all  $A_j^\ell$  ( $\ell = 1, \dots, d$ ) are given by (45) with  $\lambda_j^\ell \in \mathbb{R}$ . Secondly,  $s_j = 2\tilde{s}_j$  is even and  $A_j^\ell$  is a  $\tilde{s}_j \times \tilde{s}_j$  square block matrix given by

$$A_j^\ell = \begin{pmatrix} R_2(\lambda_j^\ell, \mu_j^\ell) & A_{\ell,j}^{12} & \dots & A_{\ell,j}^{\tilde{s}_j} \\ 0 & R_2(\lambda_j^\ell, \mu_j^\ell) & \dots & A_{\ell,j}^{2\tilde{s}_j} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & R_2(\lambda_j^\ell, \mu_j^\ell) \end{pmatrix}, \quad \ell = 1, \dots, d, \quad (46)$$

where

$$R_2(\lambda, \mu) := \begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}, \quad \lambda, \mu \in \mathbb{R}, \quad (47)$$

and  $A_{\ell,j}^{rs}$  are appropriate real matrices.

Following the decomposition (45) (respectively, (46)) we define  $\tilde{\lambda}^j$  by

$$\tilde{\lambda}^k = {}^t(\lambda_1^k, \dots, \lambda_m^k) \in \mathbb{K}^m, \quad k = 1, \dots, d. \quad (48)$$

Then we assume

$$\tilde{\lambda}^1, \dots, \tilde{\lambda}^d \text{ are linearly independent in } \mathbb{K}^m. \quad (49)$$

One can easily see that (49) is invariantly defined.

By (44) we define

$$\tilde{\lambda}_j = {}^t(\lambda_j^1, \dots, \lambda_j^d) \in \mathbb{K}^d, \quad j = 1, \dots, m, \quad (50)$$

and

$$\Lambda_m := \{\vec{\lambda}_1, \dots, \vec{\lambda}_m\}. \quad (51)$$

We define the cone  $\Gamma[\Lambda_m]$  by

$$\Gamma[\Lambda_m] = \left\{ \sum_{j=1}^m t_j \vec{\lambda}_j \in \mathbb{K}^d; t_j \geq 0, j = 1, \dots, m, \sum_{j=1}^m t_j \neq 0 \right\}. \quad (52)$$

**Definition 17** A  $\mathbb{K}^d$ -action  $\rho$  is called a Poincaré morphism if there exists a basis  $\Lambda_m \subset \mathbb{K}^m$  such that  $\Gamma[\Lambda_m]$  is a proper cone in  $\mathbb{K}^m$ , namely it does not contain a straight real line. If the condition is not satisfied, then, we say that the  $\mathbb{K}^d$ -action is in a Siegel domain.

Note that the definition is invariant under the choice of the basis  $\Lambda_m$ .

**Remark 18** The geometric definition above is equivalent to the notion of Poincaré morphism given by Stolovitch (Definition 6.2.1 in [46]).

Next, we need to introduce the notion of simultaneous resonance. For  $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{K}^m$ ,  $\beta = (\beta_1, \dots, \beta_m) \in \mathbb{K}^m$ , we set  $\langle \alpha, \beta \rangle = \sum_{v=1}^m \alpha_v \beta_v$ . For a positive integer  $k$  we define  $\mathbb{Z}_+^m(k) = \{\alpha \in \mathbb{Z}_+^m; |\alpha| \geq k\}$ . Put

$$\omega_j(\alpha) = \sum_{v=1}^d |\tilde{\lambda}^v, \alpha - \lambda_j^v|, \quad j = 1, \dots, m, \quad (53)$$

$$\omega(\alpha) = \min\{\omega_1(\alpha), \dots, \omega_m(\alpha)\}. \quad (54)$$

**Definition 19** The cone  $\Lambda_m$  is called simultaneously non-resonant (or, in short  $\rho$  is simultaneously nonresonant), if

$$\omega(\alpha) \neq 0, \quad \forall \alpha \in \mathbb{Z}_+^m(2). \quad (55)$$

If (55) does not hold, then  $\Lambda_m$  is said to be simultaneously resonant.

Clearly, the simultaneously nonresonant condition (55) is invariant under a change of the basis  $\Lambda_m$ .

The next assertion provides a geometrically invariant condition guaranteeing that the simultaneous reduction to normal form does not depend on (small) nilpotent perturbation of the linear part.

**Theorem 20** Let  $\rho$  be a Poincaré morphism. Then  $\rho$  is conjugated to a polynomial action by a convergent change of variables.

**Remark 21** As a corollary of Theorem 20 for vector fields having linear parts with nontrivial Jordan blocks one obtains generalizations of results for the existence of convergent normal forms for analytic vector fields admitting symmetries cf. [5,12,13].

**Example 22** Let  $\rho$  be a  $\mathbb{R}^2$ -action in  $\mathbb{R}^n$ ,  $n \geq 4$  with  $m = 3$ . Choose a basis  $\Lambda_2$  of  $\mathbb{R}^3$  such that

$$\Lambda_2 = \{ {}^t(1, 1, \nu), {}^t(0, 1, \mu) \}, \quad \nu, \mu \in \mathbb{R}. \quad (56)$$

By (52),  $\Gamma[\Lambda_2]$  is generated by the set of vectors  $\{(1, 0), (1, 1), (\nu, \mu)\}$ . Hence the action is a Poincaré morphism if and only if these vectors generate a proper cone, namely  $(\nu, \mu)$  is not in the set  $\{(\nu, \mu) \in \mathbb{R}^2; \nu \leq \mu \leq 0\}$ . We note that the interesting case is  $\mu < \nu \leq 0$ , where every generator in (56) is in a Siegel domain.

Next, given a two-dimensional Lie algebra, choose a basis  $X_1, X_2$  with linear parts  $A_j \in GL(4; \mathbb{C})$  satisfying  $\text{spec}(A_1) = \{1, 1, \nu, \nu\}$  and  $\text{spec}(A_2) = \{0, 1, \mu, \mu\}$ , respectively, where  $\nu < \mu < 0$ ,  $(\nu, \mu) \notin \mathbb{Q}^2$ , and

$$A_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \nu & 0 \\ 0 & 0 & 0 & \nu \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \mu & \varepsilon \\ 0 & 0 & 0 & \mu \end{pmatrix}, \quad (57)$$

where  $\varepsilon \neq 0$ .

We show a refinement of the divergence result in Gevrey classes in [54] for the solution  $v$  of the overdetermined systems of linear homological equations  $L_j v := \nabla v(x) A_j x - A_j v = f^j$  ( $j = 1, 2$ ), with the compatibility conditions for the RHS (see [54] for more details).

**Theorem 23** Let  $1/2 \leq \tau_0 < \infty$ . Then there exists

$$E_0 \subset \{(\nu, \mu) \in (\mathbb{R} \setminus \mathbb{Q})^2; \nu < \mu < 0, \\ v \text{ does not satisfy the Bruno condition}\}$$

with the density of continuum such that for every  $(\nu, \mu) \in E_0$ , there exists analytic  $f = {}^t(f^1, f^2) \in (\mathbb{C}_2^4\{x\})^2$ , satisfying the compatibility condition for the overdetermined system and such that the unique formal solution  $v(x)$  is not in  $\bigcup_{1 \leq \sigma < 2+\tau} G^\sigma(\mathbb{C}^4)$ . Moreover, for every analytic  $f$  we can find  $C > 0$  such that the unique formal solution satisfies the anisotropic Gevrey estimates

$$|v_\alpha| \leq C^{|\alpha|+1} (\alpha_3 + \alpha_4)^{(1+\tau)\alpha_4}, \quad \alpha \in \mathbb{Z}_+^4(2), \quad (58)$$

### Analytic Maps near a Fixed Point in the Presence of Jordan Blocks

As for the real analytic local diffeomorphisms preserving the origin in  $\mathbb{R}^n$ , one has in fact to deal necessarily with hyperbolic maps (cf. [4]).

Recall first the complex analytic case cf. [3,4]. Let  $\Phi(x)$  be a biholomorphic map of  $\mathbb{C}^n$  preserving the origin and  $\Phi'(0)$  the Jacobian matrix at the origin. Denote by  $\text{spec}(\Phi'(0)) = \{\lambda_1, \dots, \lambda_n\}$  the spectrum of  $\Phi'(0)$ . Clearly  $\lambda_j \neq 0$  for all  $j = 1, \dots, n$ . We define the set of all resonance multiindices of  $\Phi$  (actually it depends only on  $\text{spec}(\Phi'(0))$ ) as follows:

$$\text{Res}[\lambda_1, \dots, \lambda_n] = \bigcup_{j=1}^n \text{Res}^j[\lambda_1, \dots, \lambda_n] \quad (59)$$

$$\text{Res}^j[\lambda_1, \dots, \lambda_n] = \{\alpha \in \mathbb{Z}_+^n(2) : \lambda^\alpha - \lambda_j = 0\};$$

where  $\mathbb{Z}_+^n(k) := \{\alpha \in \mathbb{Z}_+^n : |\alpha| \geq k\}$ ,  $\lambda = (\lambda_1, \dots, \lambda_n)$  and  $\lambda^\alpha = \lambda_1^{\alpha_1} \dots \lambda_n^{\alpha_n}$  for  $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{Z}_+^n$ .

Given  $\Lambda \in GL(n, \mathbb{C})$  define  $\mathcal{O}[\Lambda]$  as the germ of all local complex analytic diffeomorphisms (biholomorphic maps)  $\Phi(x)$  of  $\mathbb{C}^n$  with 0 as a fixed point and such that  $\Phi'(0)$  is in the  $GL(n, \mathbb{C})$ -conjugacy class of  $\Lambda$ . We stress that if  $\Lambda$  is diagonalizable (semisimple) then the germ is determined by the spectrum of  $\Lambda$ , i. e., by  $n$  nonzero complex numbers  $\lambda_1, \dots, \lambda_n$ .

We will say that  $\Phi(x)$  is formally linearizable if there exists a formal series  $u(x) = x + \sum_{\alpha \in \mathbb{Z}_+^n(2)} u_\alpha x^\alpha$  such that

$$u^{-1} \circ \Phi \circ u = \Phi'(0) \quad \text{formally in } (\mathbb{C}[x])^n, \quad (60)$$

where  $\mathbb{C}[x]$  stands for the set of all formal power series with complex coefficients. It is well known (the Poincaré–Dulac theorem, cf. [3]) that under the nonresonance hypothesis  $\text{Res}[\lambda_1, \dots, \lambda_n] = \emptyset$ , i. e.,

$$\lambda^\alpha - \lambda_j \neq 0, \quad \alpha \in \mathbb{Z}_+^n(2), \quad j = 1, \dots, n, \quad (61)$$

$\Phi$  is formally linearizable. In fact, (61) is a necessary and sufficient condition in order that every holomorphic  $\Phi(x)$  with  $\text{spec}\Phi'(0) = \{\lambda_1, \dots, \lambda_n\}$  is formally linearizable. We refer to Gramchev and Walcher [25] for formal and algebraic aspects of normal forms of maps.

The formal solution of (60) involves expressions of the form  $(\lambda^\alpha - \lambda_j)^{-1}$ ; so when  $\inf_\alpha |\lambda^\alpha - \lambda_j| = 0$  for some  $j \in \{1, \dots, n\}$ , the convergence of  $u$  becomes a subtle question.

One of the main problems, starting from the pioneering work of Siegel [44], has been (and still is) to find general conditions which guarantee that  $u(x)$  converges,

i. e., that  $\Phi(x)$  is linearizable. We recall the state of the art of this subject. If the linear part  $\Phi'(0)$  is semisimple (i. e.,  $\Phi'(0)$  has no nontrivial Jordan blocks), then it is well known that for convergence we need arithmetic (Diophantine) conditions on

$$\omega(m) := \min_{2 \leq |\alpha| \leq m, 1 \leq j \leq n} |\lambda^\alpha - \lambda_j|, \quad m \in \mathbb{Z}_+(2). \quad (62)$$

We refer for the history and references to the survey paper by Herman [28]. The best condition that implies linearizability for all maps with a given semisimple linear part is due to Bruno [9], and can be expressed as (following Herman, p. 143 in [28])

$$\sum_{k=1}^{\infty} 2^{-k} \ln(\omega^{-1}(2^{k+1})) < \infty. \quad (63)$$

If one assumes the Poincaré condition

$$\max_{1 \leq j \leq n} |\lambda_j| < 1 \quad \text{or} \quad \min_{1 \leq j \leq n} |\lambda_j| > 1, \quad (64)$$

then  $\Phi$  is always analytically equivalent to its linear part  $\Phi'(0)$ , provided the nonresonance hypothesis holds. More generally, (64) implies that there are finitely many resonances and, according to the Poincaré–Dulac theorem, we can find a local biholomorphic change of the variables  $u$  bringing  $\Phi$  to normal form

$$(u^{-1} \circ \Phi \circ u)(x) = \Phi'(0)x + P_{\text{res}}(x), \quad (65)$$

where the remainder  $P_{\text{res}}(x)$  is a polynomial map containing only resonant terms.

Little is known about the (non)linearizability of  $\Phi$  in the analytic category if the Poincaré condition doesn't hold and the matrix  $\Phi'(0)$  has at least one nontrivial Jordan block. When  $n = 2$  and  $A$  has a double eigenvalue  $\lambda_1 = \lambda_2$ ,  $|\lambda_1| = 1$  and a nontrivial Jordan block, then in general  $\Phi$  is not linearizable. This result is contained in Proposition 3, p. 143 in [28], which is a consequence of results of Ilyashenko [32] and Yoccoz (the latter proved in 1978; for published proofs we refer to Appendix, pp. 86–87 in [52]). It is not difficult to extend this negative result to  $\mathbb{C}^n$ ,  $n \geq 3$ .

In a recent paper of DeLatte and Gramchev [17] biholomorphic maps in  $\mathbb{C}^n$ ,  $n = 3$  and  $n = 4$  having a single nontrivial Jordan block of the linear part have been studied. We mention also the paper of Abate [1], where nondiagonalizable discrete holomorphic dynamical systems have been investigated using geometrical tools.

One observes that in the real case, as the matrix  $\Phi'(0)$  is real, the nonresonance condition excludes eigenvalues on the unit circle, i. e., the map is hyperbolic.

As a corollary from results of Delatte and Gramchev [17] on nonsolvability of the LHE in the presence of Jordan blocks in the linear parts of biholomorphic maps preserving the origin of  $\mathbb{C}^3$  and the fundamental results of Pérez Marco [39] one gets readily the following assertion for hyperbolic analytic maps preserving the origin in  $\mathbb{R}^n$  and having nondiagonalizable linear parts at the origin.

**Theorem 24** *Let  $A$  be a real  $n \times n$  matrix such that its eigenvalues are nonresonant and lie outside the unit circle in  $\mathbb{C}$ , and  $A$  is neither expansive nor contractive (i.e., the Poincaré condition (64) is not satisfied). Let  $R(x) = (R_1(x), \dots, R(x))$  be a real valued analytic function near the origin, at least quadratic near  $x = 0$ , i.e.  $R(0) = 0$ ,  $\nabla R(0) = 0$ . Then there exists  $\eta_0 > 0$  such that for almost all  $\eta \in ]0, \eta_0]$  in the sense of the Lebesgue measure, the local analytic diffeomorphism defined by*

$$\Phi(x) = Ax + \eta R(x) \quad (66)$$

*is not linearizable by convergent transformations.*

### Weakly Hyperbolic Systems and Nilpotent Perturbations

The presence of nondiagonalizable matrices appear as a challenging problem in the framework of the well-posedness of the Cauchy problem for evolution partial differential equations. In order to illustrate the main features we focus on linear hyperbolic systems with constant coefficients in space dimension one

$$\partial_t u + A \partial_x u + Bu = 0, \quad t > 0, \quad x \in \mathbb{R} \quad (67)$$

$$u(0, x) = u^0(x) \quad (68)$$

where  $A$  and  $B$  are real  $m \times m$  matrices,  $u = (u_1, \dots, u_m)$  stands for a vector-valued smooth function. One is interested in the  $C^\infty$  well-posedness of the Cauchy problem (67), (68), namely, for every initial data  $u^0 \in (C^\infty(\mathbb{R}))^m$  there exists a unique solution  $u \in C^\infty(\mathbb{R}^2)$  of (67), (68).

We recall that the system is hyperbolic if the characteristic equation

$$|A - \lambda I| = 0 \quad (69)$$

has  $m$  real solutions  $\lambda_1, \dots, \lambda_m$ . If the roots are distinct, the system is called strictly hyperbolic. The strict hyperbolicity implies that  $A$  is diagonalizable, which in turn implies that, after a linear change of variables, one can assume  $A$  to be diagonal and reduce essentially the problem to  $m$  first-order scalar equations. In that case the Cauchy problem is well-posed in the classical sense, namely for every

smooth initial data  $u^0$  (it is enough  $u^0 \in C^1(\mathbb{R})$ ) there exists a unique smooth solution  $u(t, x)$  to the Cauchy problem. In fact, it is enough to require that  $A$  is semisimple (i.e., allowing multiple eigenvalues but excluding nilpotent parts) in order to have well-posedness (e.g., cf. the book of Taylor [48] and the references therein for more details on strictly hyperbolic systems with variable coefficients, and more general set-up in the framework of pseudodifferential operators).

If the system is hyperbolic but not strictly hyperbolic (it is called also weakly hyperbolic), it means that the matrix  $A$  has multiple eigenvalues. Here the influence of the Jordan block structure is decisive and one has non existence theorems in the  $C^\infty$  category unless one imposes additional restrictions on the lower-order term  $B$ . In many applications one encounters weakly hyperbolic systems. One example is in the so-called water waves problem, concerning the motion of the free surface of a body of an incompressible irrotational fluid under the influence of gravity (see [15] and the references therein), where for the linearized system one encounters a  $2 \times 2$  matrix of the type

$$A = \begin{pmatrix} c & \kappa \\ 0 & c \end{pmatrix},$$

where  $c$  stands for the velocity, and  $\kappa$  is a nonzero real number.

The assertions, even in the seemingly simple model cases, are not easy to state in simple terms. The first results on such systems are due to Kajitani [34]. The classification problem was completely settled for the so-called hyperbolic systems of constant multiplicities by Vaillant [50] using subtle linear algebra arguments with multiparameter dependence if the space dimension is greater than 1.

We illustrate the assertions for  $m = 2$  and  $m = 3$ , where the influence of the nilpotent part is somewhat easier to describe in details. In what follows we rewrite the results in [34, 50] by means of the Jordan block structures.

The case  $m = 2$  is easy.

**Proposition 25** *Let*

$$A = \begin{pmatrix} \lambda_0 & 1 \\ 0 & \lambda_0 \end{pmatrix}, \quad (70)$$

*for some  $\lambda_0 \in \mathbb{R}$ . Then the Cauchy problem is  $C^\infty$  well posed iff  $b_{21} = 0$ , with*

$$B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

Let the  $3 \times 3$  real matrix  $A$  have a triple eigenvalue  $\lambda_0$  and be not semisimple. Then we are reduced (modulo conjugation with an invertible matrix) to two possibilities: either

the JCF of  $A$  is a maximal Jordan block

$$\begin{pmatrix} \lambda_0 & 1 & 0 \\ 0 & \lambda_0 & 1 \\ 0 & 0 & \lambda_0 \end{pmatrix}, \quad (71)$$

or the degenerate case of one  $2 \times 2$  and one  $1 \times 1$  elementary Jordan blocks

$$A = \begin{pmatrix} \lambda_0 & 1 & 0 \\ 0 & \lambda_0 & 0 \\ 0 & 0 & \lambda_0 \end{pmatrix}, \quad (72)$$

**Proposition 26** *Let  $n = 3$ . Then the following assertions hold:*

(i) *let  $A$  be defined by (71). Then the Cauchy problem is well-posed in  $C^\infty$  iff the entries of the matrix*

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

*satisfy the identities*

$$b_{31} = b_{21} + b_{32} = b_{11} - b_{13} = 0; \quad (73)$$

(ii) *suppose that  $A$  is given by (72). The well-posedness in  $C^\infty$  holds iff*

$$b_{21} = b_{21}b_{32} = 0. \quad (74)$$

In an interesting work, Petkov [38], using real Jordan block structures depending on parameters and reduction to normal forms of matrices depending on parameters (cf. [2]), derived canonical microlocal forms for the full symbol of a pseudodifferential system with real characteristics of constant multiplicity and applies them to study the propagation of singularities of solutions of certain systems.

More generally, the Cauchy problem for hyperbolic systems with multiple characteristics have been studied by various authors where (implicitly) conditions on the nilpotent perturbations and the lower order term are imposed (e. g., cf. [8,34,51], and the references therein).

It would be interesting to write down such conditions in terms of conditions on the nilpotent perturbations on the principal part and the lower-order terms.

Finally, we mention also the work of Ghedamsi, Gourdin, Mechab, and Takeuchi [22], concerned with the Cauchy problem for Schrödinger-type systems with characteristic roots of multiplicity two admitting nontrivial Jordan blocks.

## Bibliography

- Abate M (2000) Diagonalization of nondiagonalizable discrete holomorphic dynamical systems. *Amer J Math* 122:757–781
- Arnold VI (1971) Matrices depending on parameters. *Uspekhi Mat Nauk* 26:101–114 (in Russian); *Russ Math Surv* 26:29–43
- Arnold VI (1983) Geometrical methods in the theory of ordinary differential equations. Springer, New York
- Arnold VI, Ilyashenko YU (1988) In: Anosov DV, Arnold VI (eds) *Encyclopedia of Math Sci*, vol 1. Dynamical Systems I. Springer, New York, pp 1–155
- Bambusi D, Cicogna G, Gaeta G, Marmo G (1998) Normal forms, symmetry and linearization of dynamical systems. *J Phys A* 31:5065–5082
- Belitskii GR (1978) Equivalence and normal forms of germs of smooth mappings. *Uspekhi Mat Nauk* 33:95–155, 263 (in Russian); *Russ Math Surv* 33:107–177
- Belitskii GR (1979) Normal forms, invariants, and local mappings. *Naukova Dumka*, Kiev (in Russian)
- Bove A, Nishitani T (2003) Necessary conditions for hyperbolic systems. *Il. Jpn J Math (NS)* 29:357–388
- Bruno AD (1971) The analytic form of differential equations. *Tr Mosk Mat O-va* 25:119–262; (1972) 26:199–239 (in Russian); See also (1971) *Trans Mosc Math Soc* 25:131–288; (1972) 26:199–239
- Bruno AD, Walcher S (1994) Symmetries and convergence of normalizing transformations. *J Math Anal Appl* 183:571–576
- Chen KT (1965) Diffeomorphisms:  $C^\infty$ -realizations of formal properties. *Amer J Math* 87:140–157
- Cicogna G, Gaeta G (1999) Symmetry and perturbation theory in nonlinear dynamics. *Lecture Notes in Physics. New Series m: Monographs*, vol 57. Springer, Berlin
- Cicogna G, Walcher S (2002) Convergence of normal form transformations: the role of symmetries. (English summary) *Symmetry and perturbation theory. Acta Appl Math* 70:95–111
- Coddington EA, Levinson N (1955) *Theory of ordinary differential equations*. McGraw-Hill, New York
- Craig W (1987) Nonstrictly hyperbolic nonlinear systems. *Math Ann* 277:213–232
- Cushman R, Sanders JA (1990) A survey of invariant theory applied to normal forms of vector fields with nilpotent linear part. In: Stanton D (ed) *Invariant theory and tableaux*. IMA Vol Math Appl, vol 19. Springer, New York, pp 82–106
- DeLatte D, Gramchev T (2002) Biholomorphic maps with linear parts having Jordan blocks: Linearization and resonance type phenomena. *Math Phys Electron J* 8(2):1–27
- Dumortier F, Roussarie R (1980) Smooth linearization of germs of  $R^2$ -actions and holomorphic vector fields. *Ann Inst Fourier Grenoble* 30:31–64
- Gaeta G, Walcher S (2005) Dimension increase and splitting for Poincaré-Dulac normal forms. *J Nonlinear Math Phys* 12(1):327–342
- Gaeta G, Walcher S (2006) Embedding and splitting ordinary differential equations in normal form. *J Differ Equ* 224:98–119
- Gantmacher FR (1959) *The theory of matrices*, vols 1, 2. Chelsea, New York
- Ghedamsi M, Gourdin D, Mechab M, Takeuchi J (2002) Équations et systèmes du type de Schrödinger à racines caractéristiques de multiplicité deux. *Bull Soc Roy Sci Liège* 71:169–187



23. Gramchev T (2002) On the linearization of holomorphic vector fields in the Siegel Domain with linear parts having nontrivial Jordan blocks. In: Abenda S, Gaeta G, Walcher S (eds) *Symmetry and perturbation theory*, Cala Gonone, 16–22 May 2002. World Sci. Publ., River Edge, pp 106–115
24. Gramchev T, Tolis E (2006) Solvability of systems of singular partial differential equations in function spaces. *Integral Transform Spec Funct* 17:231–237
25. Gramchev T, Walcher S (2005) Normal forms of maps: formal and algebraic aspects. *Acta Appl Math* 85:123–146
26. Gramchev T, Yoshino M (2007) Normal forms for commuting vector fields near a common fixed point. In: Gaeta G, Vitolo R, Walcher S (eds) *Symmetry and Perturbation theory*, Oltranto, 2–9 June 2007. World Sci Publ, River Edge, pp 203–217
27. Hasselblatt B, Katok A (2003) *A first course in dynamics: with a panorama of recent developments*. Cambridge University Press, Cambridge
28. Herman M (1987) Recent results and some open questions on Siegel's linearization theorem of germs of complex analytic diffeomorphisms of  $C^n$  near a fixed point. VIIIth international congress on mathematical physics, Marseille 1986. World Sci Publ, Singapore, pp 138–184
29. Hibino M (1999) Divergence property of formal solutions for first order linear partial differential equations. *Publ Res Inst Math Sci* 35:893–919
30. Hibino M (2003) Borel summability of divergent solutions for singular first order linear partial differential equations with polynomial coefficients. *J Math Sci Univ Tokyo* 10:279–309
31. Hibino M (2006) Formal Gevrey theory for singular first order quasi-linear partial differential equations. *Publ Res Inst Math Sci* 42:933–985
32. Il'yashenko Y (1979) Divergence of series reducing an analytic differential equation to linear form at a singular point. *Funct Anal Appl* 13:227–229
33. Iooss G, Lombardi E (2005) Polynomial normal forms with exponentially small remainder for vector fields. *J Differ Equ* 212:1–61
34. Kajitani K (1979) Cauchy problem for non-strictly hyperbolic systems. *Publ Res Inst Math* 15:519–550
35. Katok A, Katok S (1995) Higher cohomology for Abelian groups of toral automorphisms. *Ergod Theory Dyn Syst* 15:569–592
36. Murdock J (2002) On the structure of nilpotent normal form modules. *J Differ Equ* 180:198–237
37. Murdock J, Sanders JA (2007) A new transvectant algorithm for nilpotent normal forms. *J Differ Equ* 238:234–256
38. Petkov VM (1979) Microlocal forms for hyperbolic systems. *Math Nachr* 93:117–131
39. Pérez Marco R (2001) Total convergence or small divergence in small divisors. *Commun Math Phys* 223:451–464
40. Ramis J-P (1984) Théorèmes d'indices Gevrey pour les équations différentielles ordinaires. *Mem Amer Math Soc* 48:296
41. Rodino L (1993) *Linear partial differential operators in Gevrey spaces*. World Science, Singapore
42. Sanders JA (2005) Normal form in filtered Lie algebra representations. *Acta Appl Math* 87:165–189
43. Sanders JA, Verhulst F, Murdock J (2007) *Averaging methods in nonlinear dynamical systems*, 2nd edn. Applied Mathematical Sciences, vol 59. Springer, New York
44. Siegel CL (1942) Iteration of analytic functions. *Ann Math* 43:607–614
45. Sternberg S (1958) The structure of local homeomorphisms. II, III. *Amer J Math* 80:623–632, 81:578–604
46. Stolovitch L (2000) Singular complete integrability. *Publ Math IHES* 91:134–210
47. Stolovitch L (2005) Normalisation holomorphe d'algèbres de type Cartan de champs de vecteurs holomorphes singuliers. *Ann Math* 161:589–612
48. Taylor M (1981) *Pseudodifferential operators*. Princeton Mathematical Series, vol 34. Princeton University Press, Princeton
49. Thiffeault J-L, Morison PJ (2000) Classification and Casimir invariants of Lie–Poisson brackets. *Physica D* 136:205–244
50. Vaillant J (1999) Invariants des systèmes d'opérateurs différentiels et sommes formelles asymptotiques. *Jpn J Math (NS)* 25:1–153
51. Yamahara H (2000) Cauchy problem for hyperbolic systems in Gevrey class. A note on Gevrey indices. *Ann Fac Sci Toulouse Math* 19:147–160
52. Yoccoz J-C (1995) A remark on Siegel's theorem for nondiagonalizable linear part. Manuscript, 1978; See also Théorème de Siegel, nombres de Bruno e polynômes quadratic. *Astérisque* 231:3–88
53. Yoshino M (1999) Simultaneous normal forms of commuting maps and vector fields. In: Degasperis A, Gaeta G (eds) *Symmetry and perturbation theory SPT 98*, Rome, 16–22 December 1998. World Scientific, Singapore, pp 287–294
54. Yoshino M, Gramchev T (2008) Simultaneous reduction to normal forms of commuting singular vector fields with linear parts having Jordan blocks. *Ann Inst Fourier (Grenoble)* 58:263–297
55. Zung NT (2002) Convergence versus integrability in Poincaré–Dulac normal form. *Math Res Lett* 9:217–228

---

## Perturbation Theory

GIOVANNI GALLAVOTTI

Dipartimento di Fisica and I.N.F.N., sezione Roma-I,  
Università di Roma I “La Sapienza”, Roma, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Poincaré's Theorem and Quanta](#)

[Mathematics and Physics. Renormalization](#)

[Need of Convergence Proofs](#)

[Multiscale Analysis](#)

[A Paradigmatic Example of PT Problem](#)

[Lindstedt Series](#)

[Convergence. Scales. Multiscale Analysis](#)

[Non Convergent Cases](#)

[Conclusion and Outlook](#)

[Future Directions](#)

[Bibliography](#)

## Glossary

**Formal power series** a power series, giving the value of a function  $f(\varepsilon)$  of a parameter  $\varepsilon$ , that is derived assuming that  $f$  is analytic in  $\varepsilon$ .

**Renormalization group** method for multiscale analysis and resummation of formal power series. Usually applied to define a systematic collection of terms to organize a formal power series into a convergent one.

**Lindstedt series** an algorithm to develop formal power series for computing the parametric equations of invariant tori in systems close to integrable.

**Multiscale problem** any problem in which an infinite number of scales play a role.

## Definition of the Subject

**Perturbation Theory:** Computation of a quantity depending on a parameter  $\varepsilon$  starting from the knowledge of its value for  $\varepsilon = 0$  by deriving a power series expansion in  $\varepsilon$ , under the assumption of its existence, and if possible discussing the interpretation of the series. Perturbation theory is very often the only way to get a glimpse of the properties of systems whose equations cannot be “explicitly solved” in computable form.

The importance of Perturbation Theory is witnessed by its applications in Astronomy, where it led not only to the discovery of new planets (Neptune) but also to the discovery of Chaotic motions, with the completion of the Copernican revolution and the full understanding of the role of Aristotelian Physics formalized into uniform rotations of deferents and epicycles (today Fourier representation of quasi periodic motions). It also played an essential role in the development of Quantum Mechanics and the understanding of the periodic table. The successes of Quantum Field Theory in Electrodynamics first, then in Strong interactions and finally in the unification of the elementary forces (strong, electromagnetic, and weak) are also due to perturbation theory, which has also been essential in the theoretical understanding of the critical point universality. The latter two themes concern the new methods that have been developed in the last fifty years, marking a kind of new era for perturbation theory; namely dealing with singular problems, via the techniques called, in Physics, “Renormalization Group” and, in Mathematics, “Multiscale Analysis”.

## Introduction

Perturbation theory, henceforth PT, arises when the value of a function of interest is associated with a problem depending on a parameter, here called  $\varepsilon$ . The value has to be

a simple, or at least explicit and rigorous, computation for  $\varepsilon = 0$  while its computation for  $\varepsilon \neq 0$ , small, is attempted by expressing it as the sum of a power series in  $\varepsilon$  which will be called here the “solution”.

It is important to say since the beginning that a real PT solution of a problem involves two distinct steps: the first is to show that assuming that there is a convergent power series solving the problem then the *coefficients* of the  $n$ th power of  $\varepsilon$  exist and can be computed via finite computation. The resulting series will be called *formal solution* or *formal series* for the problem. The second step, that will be called *convergence theory*, is to prove that the formal series converges for  $\varepsilon$  small enough, or at least find a “summation rule” that gives a meaning to the formal series thus providing a real solution to the problem. None of the two problems is trivial, in the interesting cases, although the second is certainly the key and a difficult one.

Once Newton’s law of universal gravitation was established it became necessary to develop methods to find its implications. Laplace’s “*Mécanique Céleste*” [19], provided a detailed and meticulous exposition of a general method that has become a classic, if not the first, example of perturbation theory, quite different from the parallel analysis of Gauss which can be more appropriately considered a “non perturbative” development.

Since Laplace one can say that many applications along his lines followed. In the XIX century wide attention was dedicated to extend Laplace’s work to cover various astronomical problems: tables of the coefficients were dressed and published, and algorithms for their construction were devised, and planets were discovered (Neptune, 1846). Well known is the “Lindstedt algorithm” for the computation of the  $n$ th order coefficients of the PT series for the non resonant quasi periodic motions. The algorithm provides a power series representation for the quasi periodic motions with non resonant frequencies which is extremely simple: however it represents the  $n$ th coefficient as a sum of many terms, some of size of the order of a power on  $n!$ . Which of course is a serious problem for the convergence.

It became a central issue, known as the “small denominators problem” after Poincaré’s deep critique of the PT method, generated by his analysis of the three-body problem. It led to his “non-integrability theorem” about the generic nonexistence of convergent power series in the perturbation parameter  $\varepsilon$  whose sum would be a constant of motion for a Hamiltonian  $H_\varepsilon$ , member of a family of Hamiltonians parametrized by  $\varepsilon$  and reducing to an integrable system for  $\varepsilon = 0$ . The theorem suggested (to some) that even the PT series of Lindstedt (to which Poincaré’s theorem does not apply) could be meaningless even though formally well defined [23].

A posteriori, it should be recognized that PT was involved also in the early developments of Statistical Mechanics in the XIX century: the virial theorem application to obtain the Van der Waals equation of state can be considered a first-order calculation in PT (although this became clear only a century later with the identification of  $\varepsilon$  as the inverse of the space dimension).

### Poincaré's Theorem and Quanta

With Poincaré begins a new phase: the question of convergence of series in  $\varepsilon$  becomes a central one in the Mathematics literature. Much less, however, in the Physics literature where the new discoveries in the atomic phenomena attracted the attention. It seems that in the Physics research it was taken for granted that convergence was not an issue: atomic spectra were studied via PT and early authoritative warnings were simply disregarded (for instance, explicit by Einstein, in [1], and clear, in [3], but "timid" being too far against the mainstream, for his young age). In this way quantum theory could grow from the original formulations of Bohr, Sommerfeld, Ehrenfest relying on PT to the final formulations of Heisenberg and Schrödinger quite far from it. Nevertheless, the triumph of quantum theory was quite substantially based on the technical development and refinement of the methods of formal PT: the calculation of the Compton scattering, the Lamb shift, Fermi's weak interactions model and other spectacular successes came in spite of the parallel recognition that some of the series that were being laboriously computed not only could not possibly be convergent but their very existence, to all orders  $n$ , was in doubt.

The later Feynman graphs representation of PT was a great new tool which superseded and improved earlier graphical representations of the calculations. Its simplicity allowed a careful analysis and understanding of cases in which even *formal* PT seemed puzzlingly failing.

Renormalization theory was developed to show that the convergence problems that seemed to plague even the computation of the individual coefficients of the series, hence the formal PT series at fixed order, were, in reality, often absent, in great generality, as suspected by the earlier treatments of special (important) cases, like the higher-order evaluations of the Compton scattering, and other quantum electrodynamics cross sections or anomalous characteristic constants (e. g. the magnetic moment of the muon).

### Mathematics and Physics. Renormalization

In 1943 the first important result on the convergence of the series of the Lindstedt kind was obtained by Siegel [25]:

a formal PT series, of interest in the theory of complex maps, was shown to be convergent. Siegel's work was certainly a stimulus for the later work of Kolmogorov who solved [18], a problem that had been considered not soluble by many: to find the convergence conditions and the convergence proof of the Lindstedt series for the quasi periodic motions of a generic analytic Hamiltonian system, in spite of Poincaré's theorem and actually avoiding contradiction with it. Thus, showing the soundness of the comments about the unsatisfactory aspects of Poincaré's analysis that had been raised almost immediately by Weierstrass, Hadamard and others.

In 1956 not only Kolmogorov theorem appeared but also convergence of another well known and widely used formal series, the virial series, was achieved in an unnoticed work by Morrey [21], and independently rediscovered in the early 1960's.

At this time it seems that all series with well-defined terms were thought to be either convergent or at least asymptotic: for most Physicists convergence or asymptoticity were considered of little interest and matters to be left to Mathematicians.

However, with the understanding of the formal aspects of renormalization theory the interest in the convergence properties of the formal PT series once again became the center of attention.

On the one hand mathematical proofs of the existence of the PT series, for interesting quantum fields models, to all orders were investigated settling the question once and for all (Hepp's theorem [16]); on the other hand it was obvious that even if convergent (like in the virial or Meyer expansions, or in the Kolmogorov theory) it was well understood that the radius of convergence would not be large enough to cover all the physically interesting cases. The sum of the series would in general become singular in the sense of analytic functions and, even if admitting analytic continuation beyond the radius of convergence, a singularity in  $\varepsilon$  would be eventually hit. The singularity was supposed to correspond to very important phenomena, like the critical point in statistical mechanics or the onset of chaotic motions (already foreseen by Poincaré in connection with his non convergence theorem). Thus, research developed in two direction.

The first aimed at understanding the nature of the singularities from the formal series coefficients: in the 1960s many works achieved the understanding of the scaling laws (i. e. some properties of the divergences appearing at the singularities of the PT series or of its analytic continuation, for instance in the work of M. Fisher, Kadanoff, Widom and may others). This led to trying to find *resummations*, i. e. to collect terms of the formal series to trans-

form them into *convergent series* in terms of *new parameters*, the *running couplings*.

The latter would be singular functions of the original  $\varepsilon$  thus possibly reducing the study of the singularity to the singularities of the running couplings. The latter could be studied by independent methods, typically by studying the iterations of an auxiliary dynamical system (called the *beta function flow*). This was the *approach* or *renormalization group method* of Wilson [10,29].

The second direction was dedicated to finding out the real meaning of the PT series in the cases in which convergence was doubtful or a priori excluded: in fact already Landau had advanced the idea that the series could be just illusions in important problems like the paradigmatic quantum field theory of a scalar field or the fundamental quantum electrodynamics [4,15].

In a rigorous treatment the function that the series were supposed to represent would be in fact a trivial function with a dependence on  $\varepsilon$  unrelated to the coefficients of the well defined and non trivial but formal series. It was therefore important to show that there were at least cases in which the perturbation series of a nontrivial problem had a meaning determined by its coefficients. This was studied in the scalar model of quantum field theory and a proof of “non triviality” was achieved after the ground-breaking work of Nelson on two-dimensional models [22,26]: soon followed by similar results in two dimensions and the difficult extension to three-dimensional models by Glimm and Jaffe [14], and generating many works and results on the subject which took the name of “constructive field theory” [6].

But Landau’s *triviality conjecture* was actually dealing with the “real problem”, i. e. the 4-dimensional quantum fields. The conjecture remains such at the moment, in spite of very intensive work and attempts at its proof. The problem had relevance because it could have meant that not only the simple scalar models of constructive field theory were trivial but also the QED series which had received strong experimental support with the correct prediction of fine structure phenomena could be illusions, in spite of their well-defined PT series: which would remain as mirages of a non existing reality.

The work of Wilson made clear that the “triviality conjecture” of Landau could be applied only to theories which, after the mentioned resummations, would be controlled by a beta function flow that could not be studied perturbatively, and introduced the new notion of *asymptotic freedom*. This is a property of the beta function flow, implying that the running couplings are bounded and small so that the resummed series are more likely to have a meaning [29].

This work revived the interest in PT for quantum fields with attention devoted to new models that had been believed to be non renormalizable. Once more the apparently preliminary problem of developing a formal PT series played a key role: and it was discovered that many Yang–Mills quantum field theories were in fact renormalizable in the ultraviolet region [27,28], and an exciting period followed with attempts at using Wilson’s methods to give a meaning to the Yang–Mills theory with the hope of building a theory of the strong interactions. Thus, it was discovered that several Yang–Mills theories were asymptotically free as a consequence of the high symmetry of the model, proving that what seemed to be strong evidence that no renormalizable model would have asymptotic freedom was an ill-founded belief (that in a sense slowed down the process of understanding, and not only of the strong interactions).

Suddenly understanding the strong interactions, until then considered an impossible problem became possible [15], as solutions could be written and *effectively computed* in terms of PT which, although not proved to be convergent or asymptotic (still an open problem in dimension  $d = 4$ ) were immune to the argument of Landau. The impact of the new developments led a little later to the unification of all interactions into the *standard model* for the theory of elementary particles (including the electromagnetic and weak interactions). The standard model was shown to be asymptotically free *even in the presence of symmetry breaking*, at least if a few other interactions in the model (for instance the Higgs particle self interaction) were treated heuristically while waiting for the discovery of the “Higgs particle” and for a better understanding of the structure of the elementary particles at length scales intermediate between the Fermi scale ( $\sim 10^{-15}$  cm (the weak interactions scale)) and the Planck scale (the gravitational interaction scale, 15 orders of magnitude below).

Given that the very discovery of renormalizability of Yang–Mills fields and the birth of a strong interactions theory had been firmly grounded on experimental results [15], the latter “missing step” was, and still is, considered an acceptable gap.

## Need of Convergence Proofs

The story of the standard model is paradigmatic of the power of PT: it should convince anyone that the analysis of formal series, including their representation by diagrams, which plays an essential part, is to be taken seriously. PT is certainly responsible for the revival and solution of problems considered by many as hopeless.



In a sense PT in the elementary particles domain can only, so far, partially be considered a success. Different is the situation in the developments that followed the works of Siegel and Kolmogorov. Their relevance for Celestial Mechanics and for several problems in applied physics (particle accelerator design, nuclear fusion machines for instance) and for statistical mechanics made them too the object of a large amount of research work.

The problems are simpler to formulate and often very well posed but the possibility of existence of chaotic motions, always looming, made it imperative not to be content with heuristic analysis and imposed the quest of mathematically complete studies. The lead were the works of Siegel and Kolmogorov. They had established convergence of certain PT series, but there were other series which would certainly be not convergent even though formally well defined and the question was, therefore, which would be their meaning.

More precisely it was clear that the series could be used to find approximate solutions to the equations, representing the motion for very long times under the assumption of “small enough”  $\varepsilon$ . But this could hardly be considered an understanding of the PT series in Mechanics: the estimated values of  $\varepsilon$  would have to be too small to be of interest, with the exception of a few special cases. The real question was what could be done to give the PT series the status of exact solution.

As we shall see the problem is deeply connected with the above-mentioned asymptotic freedom: this is perhaps not surprising because the link between the two is to be found in the “multiscale analysis” problems, which in the last half century have been the core of the studies in many areas of Analysis and in Physics, when theoretical developments and experimental techniques became finer and able to explore nature at smaller and smaller scales.

## Multiscale Analysis

To illustrate the multiscale analysis in PT it is convenient to present it in the context of Hamiltonian mechanics, because in this field it provides us with nontrivial cases of almost complete success.

We begin by contrasting the work of Siegel and that of Kolmogorov: which are based on radically different methods. The first being much closer in spirit to the developments of renormalization theory and to the Feynman graphs.

Most interesting formal PT series have a common feature: namely their  $n$ th order coefficients are constructed as sums of many “terms” and the first attempt to a complete analysis is to recognize that their sum, which gives

the uniquely defined  $n$ th coefficient is much smaller than the sum of the absolute values of the constituent terms. This is a property usually referred to as a “cancellation” and, as a rule, it reflects some symmetry property of the problem: hence one possible approach is to look for expressions of the coefficients and for cancellations which would reduce the estimate of the  $n$ th order coefficients, very often of the order of a power of  $n!$ , to an exponential estimate  $O(\varrho^{-n})$  for some  $\varrho > 0$  yielding convergence (parenthetically in the mentioned case of Yang–Mills theories the reduction is even more dramatic as it leads from divergent expressions to finite ones, yet of order  $n!$ ).

The multiscale aspect becomes clear also in Kolmogorov’s method because the implicit functions theorem has to be applied over and over again and deals with functions implicitly defined on smaller and smaller domains [5,7]. But the method purposely avoids facing the combinatorial aspects behind the cancellations so much followed, and cherished [28], in the Physics works.

Siegel’s method was developed to study a problem in which no grouping of terms was eventually needed, even though this was by no means clear a priori [24]; and to realize that no cancellations were needed forced one to consider the problem as a multiscale one because the absence of rapid growth of the  $n$ th order coefficients became manifest after a suitable “hierarchical ordering” of the terms generating the coefficients. The approach establishes a strong connection with the Physics literature because the technique to study such cases was independently developed in quantum field theory with renormalization, as shown by Hepp in [16], relying strongly on it. This is very natural and, in case of failure, it can be improved by looking for “resummations” turning the power series into a convergent series in terms of functions of  $\varepsilon$  which are singular but controllably so. For details see below and [8].

What is “natural”, however, is a very personal notion and it is not surprising that what some consider natural is considered unnatural or clumsy or difficult (or the three qualifications together) by others.

Conflict arises when the same problem can be solved by two different “natural” methods and in the case of PT for Hamiltonian systems close to integrable ones (closeness depending on the size of a parameter  $\varepsilon$ ), the so-called “small denominators” problem, the methods of Siegel and Kolmogorov are antithetic and an example of the just mentioned dualism.

The first method, that will be called here “Siegel’s method” (see below for details), is based on a careful analysis of the structure of the various terms that occur at a given PT order achieving a proof that the  $n$ th order co-



efficient which is represented as the sum of many terms some of which *might* have size of order of a power of  $n!$  has in fact a size of  $O(\varrho^{-n})$  so that the PT series is convergent for  $|\varepsilon| < \varrho$ . Although strictly speaking the original work of Siegel does not immediately apply to the Hamiltonian Mechanics problems (see below), it can nevertheless be adapted and yields a solution, as made manifest much later in [2,8,24].

The second method, called here “Kolmogorov’s method”, instead does not consider the individual coefficients of the various orders but just regards the sum of the series as a solution of an implicit function equation (a “Hamilton–Jacobi” equation) and devises a recursive algorithm approximating the unknown sum of the PT series by functions analytic in a disk of fixed radius  $\varrho$  in the complex  $\varepsilon$ -plane [5,7].

Of course the latter approach implies that no matter how we achieve the construction of the  $n$ th order PT series coefficient there will have to be enough cancellations, if at all needed, so that it turns out bounded by  $O(\varrho^{-n})$ . And in the problem studied by Kolmogorov cancellations would be necessarily present if the  $n$ th order coefficient was represented by the sum of the terms in the Lindstedt series.

That this is not obvious is supported by the fact that it was considered an open problem, for about thirty years, to find a way to exhibit explicitly the cancellation mechanism in the Lindstedt series implied by Kolmogorov’s work. This was done by Eliasson [2], who proved that the coefficients of the PT of a given order  $n$  as expressed by the construction known as the “Lindstedt algorithm” yielded coefficients of size of  $O(\varrho^{-n})$ : his argument, however, did not identify in general which term of the Lindstedt sum for the  $n$ th order coefficient was compensated by which other term or terms. It proved that the sum had to satisfy suitable relations, which in turn implied a total size of  $O(\varrho^{-n})$ . And it took a few more years for the complete identification [8], of the rules to follow in collecting the terms of the Lindstedt series which would imply the needed cancellations.

It is interesting to remark that, aside from the example of Hamiltonian PT, multiscale problems have dominated the development of analysis and Physics in recent time: for instance they appear in harmonic analysis (Carleson, Fefferman), in PDE’s (DeGiorgi, Moser, Caffarelli–Kohn–Nirenberg), in relativistic quantum mechanics (Glimm, Jaffe, Wilson), in Hamiltonian Mechanics (Siegel, Kolmogorov, Arnold, Moser), in statistical mechanics and condensed matter (Fisher, Wilson, Widom) ... Sometimes, although not always, studied by PT techniques [10].

## A Paradigmatic Example of PT Problem

It is useful to keep in mind an example illustrating technically what it means to perform a multiscale analysis in PT. And the case of quasi periodic motions in Hamiltonian mechanics will be selected here, being perhaps the simplest.

Consider the motion of  $\ell$  unit masses on a unit circle and let  $\alpha = (\alpha_1, \dots, \alpha_\ell)$  be their positions on the circle, i. e.  $\alpha$  is a point on the torus  $\mathcal{T}^\ell = [0, 2\pi]^\ell$ . The points interact with a potential energy  $\varepsilon f(\alpha)$  where  $\varepsilon$  is a strength parameter and  $f$  is a trigonometric even polynomial, of degree  $N$ :  $f(\alpha) = \sum_{\mathbf{v} \in \mathbb{Z}^\ell, |\mathbf{v}| \leq N} f_{\mathbf{v}} e^{i\mathbf{v} \cdot \alpha}$ ,  $f_{\mathbf{v}} = f_{-\mathbf{v}} \in \mathbb{R}$ , where  $\mathbb{Z}^\ell$  denotes the lattice of the points with integer components in  $\mathbb{R}^\ell$  and  $|\mathbf{v}| = \sum_j |v_j|$ .

Let  $t \rightarrow \alpha_0 + \omega_0 t$  be the motion with initial data, at time  $t = 0$ ,  $\alpha(0) = \alpha_0$ ,  $\dot{\alpha}(0) = \omega_0$ , in which all particles rotate at constant speed with rotation velocity  $\omega_0 = (\omega_{01}, \dots, \omega_{0\ell}) \in \mathbb{R}^\ell$ . This is a solution for the equations of motion for  $\varepsilon = 0$  and it is a quasi periodic solution, i. e. each of the angles  $\alpha_j$  rotates periodically at constant speed  $\omega_{0j}$ ,  $j = 1, \dots, \ell$ .

The motion will be called *non resonant* if the components of the rotation speed  $\omega_0$  are rationally independent: this means that  $\omega_0 \cdot \mathbf{v} = 0$  with  $\mathbf{v} \in \mathbb{Z}^\ell$  is possible only if  $\mathbf{v} = \mathbf{0}$ . In this case the motion  $t \rightarrow \alpha_0 + \omega_0 t$  covers,  $\forall \alpha_0$ , densely the torus  $\mathcal{T}^\ell$  as  $t$  varies. The PT problem that we consider is to find whether there is a family of motions “of the same kind” for each  $\varepsilon$ , small enough, solving the equations of motion; more precisely whether there exists a function  $\mathbf{a}_\varepsilon(\varphi)$ ,  $\varphi \in \mathcal{T}^\ell$ , such that setting

$$\alpha(t) = \varphi + \omega_0 t + \mathbf{a}_\varepsilon(\varphi + \omega_0 t), \quad \text{for } \varphi \in \mathcal{T}^\ell \quad (1)$$

one obtains,  $\forall \varphi \in \mathcal{T}^\ell$  and for  $\varepsilon$  small enough, a solution of the equations of motion for a force  $-\varepsilon \partial_\alpha f(\alpha)$ : i. e.

$$\ddot{\alpha}(t) = -\varepsilon \partial_\alpha f(\alpha(t)). \quad (2)$$

By substitution of Eq. (1) in Eq. (2), the condition becomes  $(\omega_0 \cdot \partial_\varphi)^2 \mathbf{a}_\varepsilon(\varphi + \omega_0 t) = -\partial_\alpha f(\varphi + \omega_0 t)$ . Since  $\omega_0$  is assumed rationally independent  $\varphi + \omega_0 t$  covers densely the torus  $\mathcal{T}^\ell$  as  $t$  varies: hence the equation for  $\mathbf{a}_\varepsilon$  is

$$(\omega_0 \cdot \partial_\varphi)^2 \mathbf{a}_\varepsilon(\varphi) = -\varepsilon \partial_\alpha f(\varphi + \mathbf{a}_\varepsilon(\varphi)). \quad (3)$$

Applying PT to this equation means to look for a solution  $\mathbf{a}_\varepsilon$  which is analytic in  $\varepsilon$  small enough and in  $\varphi \in \mathcal{T}^\ell$ . In colorful language one says that the perturbation effect is slightly deforming a nonresonant torus with given frequency spectrum (i. e. given  $\omega_0$ ) on which the motion develops, without destroying it and keeping the quasi periodic motion on it with the same frequency spectrum.

### Lindstedt Series

As it follows from a very simple special case of Poincaré's work, Eq. (3) cannot be solved if also  $\omega_0$  is considered variable and the dependence on  $\varepsilon, \omega_0$  analytic. Nevertheless if  $\omega_0$  is *fixed and non resonant* and if  $\mathbf{a}_\varepsilon$  is supposed analytic in  $\varepsilon$  small enough and in  $\varphi \in \mathcal{T}^\ell$ , then there can be at most one solution to the Eq. (3) with  $\mathbf{a}_\varepsilon(\mathbf{0}) = \mathbf{0}$  (which is not a real restriction because if  $\mathbf{a}_\varepsilon(\varphi)$  is a solution also  $\mathbf{a}_\varepsilon(\varphi + \mathbf{c}_\varepsilon) + \mathbf{c}_\varepsilon$  is a solution for any constant  $\mathbf{c}_\varepsilon$ ). This so because the coefficients of the power series in  $\varepsilon$ ,  $\sum_{n=1}^{\infty} \varepsilon^n \mathbf{a}_n(\varphi)$ , are uniquely determined if the series is convergent. In fact they are trigonometric polynomials of order  $\leq nN$  which will be written as

$$\alpha_n(\varphi) = \sum_{0 < |\mathbf{v}| \leq Nn} \alpha_{n,\mathbf{v}} e^{i\mathbf{v} \cdot \varphi}. \quad (4)$$

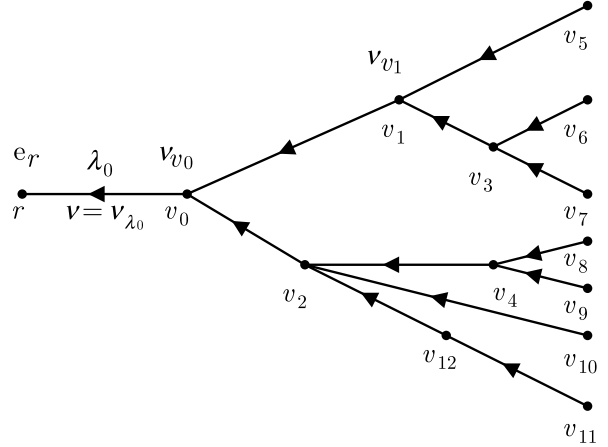
It is convenient to express them in terms of graphs. The graphs to use to express the value  $\mathbf{a}_{n,\mathbf{v}}$  are

- (i) trees with  $n$  nodes  $v_1, \dots, v_n$ ,
- (ii) one root  $r$ ,
- (iii)  $n$  lines joining pairs  $(v'_i, v_i)$  of nodes or the root and one node, *always one and not more*,  $(r, v_i)$ ;
- (iv) the lines will be different from each other and distinguished by a mark label,  $1, \dots, n$  attached to them. The connections between the nodes that the lines generate have to be *loopless*, i. e. the graph formed by the lines must be a tree.
- (v) The tree lines will be imagined oriented towards the root: hence a partial order is generated on the tree and the line joining  $v$  to  $v'$  will be denoted  $\lambda_{v',v}$  and  $v'$  will be the node closer to the root immediately following  $v$ , hence such that  $v' > v$  in the partial order of tree.

The number of such trees is large and exactly equal to  $n^{n-1}$ , as an application of Cayley's formula implies: their collection will be denoted  $T_n^0$ .

To compute  $\mathbf{a}_{n,\mathbf{v}}$  consider all trees in  $T_n^0$  and attach to each node  $v$  a vector  $\mathbf{v}_v \in \mathbb{Z}^\ell$ , called "mode label", such that  $f_{\mathbf{v}_v} \neq 0$ , hence  $|\mathbf{v}_v| \leq N$ . To the root we associate one of the coordinate unit vectors  $\mathbf{v}_r \equiv \mathbf{e}_r$ . We obtain a set  $T_n$  of *decorated trees* (with  $\leq (2N+1)^{\ell n} n^{n-1}$  elements, by the above counting analysis).

Given  $\theta \in T_n$  and  $\lambda = \lambda(v', v) \in \theta$  we define the *current* on the line  $\lambda$  to be the vector  $\mathbf{v}(\lambda) \equiv \mathbf{v}(v', v) \stackrel{\text{def}}{=} \sum_{w \leq v} \mathbf{v}_w$ : i. e. we imagine that the node vectors  $\mathbf{v}_{v_i}$  represent currents entering the node  $v_i$  and flowing towards the root. Then  $\mathbf{v}(\lambda)$  is, for each  $\lambda$ , the sum of the currents which entered all the nodes not following  $v$ , i. e. current accumulated after passing the node  $v$ .



**Perturbation Theory, Figure 1**

A tree  $\theta$  with  $m_{v_0} = 2, m_{v_1} = 2, m_{v_2} = 3, m_{v_3} = 2, m_{v_4} = 2, m_{v_{12}} = 1$  lines entering the nodes  $v_i, k = 13$ . Some labels or decorations explicitly marked (on the lines  $\lambda_0, \lambda_1$  and on the nodes  $v_1, v_2$ ); the number labels, distinguishing the branches, are not shown. The arrows represent the partial ordering on the tree

The current flowing in the root line  $\mathbf{v} = \sum_v \mathbf{v}_v$  will be denoted  $\mathbf{v}(\theta)$ .

Let  $T_n^*$  be the set trees in  $T_n$  in which *all lines* carry a *non zero current*  $\mathbf{v}(\lambda) \neq \mathbf{0}$ . A value  $\text{Val}(\theta)$  will be defined, for  $\theta \in T_n^*$ , by a product of node factors and of line factors over all nodes and

$$\text{Val}(\theta) = \frac{i(-1)^n}{n!} \prod_{v \in \theta} f_{\mathbf{v}_v} \prod_{\lambda=(v',v)} \frac{\mathbf{v}_{v'} \cdot \mathbf{v}_v}{(\omega_0 \cdot \mathbf{v}(v', v))^2}. \quad (5)$$

The coefficient  $\mathbf{a}_{n,\mathbf{v}}$  will then be

$$\mathbf{a}_{n,\mathbf{v}} = \sum_{\substack{\theta \in T_n^* \\ \mathbf{v}(\theta) = \mathbf{v}}} \text{Val}(\theta) \quad (6)$$

and, when the coefficients are imagined to be constructed in this way, the formal power series  $\sum_{n=1}^{\infty} \varepsilon^n \sum_{|\mathbf{v}| \leq Nn} \mathbf{a}_{n,\mathbf{v}}$  is called the "Lindstedt series". Eq. (5) and its graphical interpretation in Fig. 1 should be considered the "Feynman rules" and the "Feynman diagrams" of the PT for Eq. (3) [9,10].

### Convergence. Scales. Multiscale Analysis

The Lindstedt series is well defined because of the non resonance condition and the  $n$ th term is not even a sum of too many terms: if  $F \stackrel{\text{def}}{=} \max_v |f_v|$ , each of them can be bounded by  $F^n/n! \prod_{\lambda \in \theta} N^2/(\omega_0 \cdot \mathbf{v}(\lambda)^2)$ ; hence their sum can be bounded, if  $G$  is such that  $((2N+1)^{\ell n} n^{n-1} F^n)/n! \leq G^n$ , by  $G^n \prod_{\lambda \in \theta} N^2/(\omega_0 \cdot \mathbf{v}(\lambda)^2)$ .

Thus, all  $\mathbf{a}_n$  are well defined and finite *but* the problem is that  $|\mathbf{v}(\lambda)|$  can be large (up to  $Nn$  at given order  $n$ ) and therefore  $\omega_0 \cdot \mathbf{v}(\lambda)$  although never zero can become very small as  $n$  grows. For this reason the problem of convergence of the series is an example of what is called a *small denominators problem*. And it is necessary to assume more than just non resonance of  $\omega_0$  in order to solve it in the present case: a simple condition is the *Diophantine* condition, namely the existence of  $C, \tau > 0$  such that

$$|\omega_0 \cdot \mathbf{v}| \geq \frac{1}{C |\mathbf{v}|^\tau}, \quad \forall \mathbf{0} \neq \mathbf{v} \in \mathbb{Z}^\ell. \quad (7)$$

But this condition is not sufficient in an obvious way: because it only allows us to bound individual tree-values by  $n!^a$  for some  $a > 0$  related to  $\tau$ ; furthermore it is not difficult to check that there are single graphs whose value is actually of “factorial” size in  $n$ . Although non trivial to see (as mentioned above) this was only apparently so in the earlier case of Siegel’s problem but it is the new essential feature of the terms generating the  $n$ th order coefficient in Eq. (6).

A resummation is necessary to show that the tree-values can be grouped so that the sum of the values of each group can be bounded by  $\varrho^{-n}$  for some  $\varrho > 0$  and  $\forall n$ , although the group may contain (several) terms of factorial size. The terms to be grouped have to be ordered hierarchically according to the sizes of the line factors  $1/(\omega_0 \cdot \mathbf{v}(\lambda))^2$ , which are called *propagators* in [8,12].

A similar problem is met in quantum field theory where the graphs are the *Feynman graphs*: such graphs can only have a small number of lines that converge into a node but they can have loops, and to show that the perturbation series is well defined to all orders it is also necessary to collect terms hierarchically according to the propagators sizes. The systematic way was developed by Hepp [16,17], for the PT expansion of the Schwinger functions in quantum field theory of scalar fields [6]. It has been used on many occasions later and it plays a key role in the renormalization group methods in Statistical Mechanics (for instance in theory of the ground state of Fermi systems) [10,11].

However, it is in the Lindstedt series that the method is perhaps best illustrated. Essentially because it ends up in a convergence proof, while often in the field theory or statistical mechanics problems the PT series can be only proved to be well defined to all orders, but they are seldom, if ever, convergent so that one has to have recourse to other supplementary analytic means to show that the PT series are asymptotic (in the cases in which they are such).

The path of the proof is the following.

- (1) Consider only trees in which no two lines  $\lambda_+$  and  $\lambda_-$ , with  $\lambda_+$  following  $\lambda_-$  in the partial order of the tree, have the *same* current  $\mathbf{v}_0$ . In this case the maximum of the  $\prod_\lambda 1/(\omega_0 \cdot \mathbf{v}(\lambda))^2$  over all trees  $\theta \in T_n^*$  can be bounded by  $G_1^n$  for some  $G_1$ .

This is an immediate consequence and the main result in Siegel’s original work [25], which dealt with a different problem with small denominators in its formal PT solution: the coefficients of the series could also be represented by tree graphs, very similar too the ones above: but the only allowed  $\mathbf{v} \in \mathbb{Z}^\ell$  were the non zero vectors with all components  $\geq 0$ .

The latter property automatically guarantees that the graphs contain no pair of lines  $\lambda_+, \lambda_-$  following each other as above in the tree partial order and having the same current. Siegel’s proof also implies a multiscale analysis [24]: but it requires no grouping of the terms unlike the analogue Lindstedt series, Eq. (6).

- (2) Trees which contain lines  $\lambda_+$  and  $\lambda_-$ , with  $\lambda_+$  following  $\lambda_-$ , in the partial order of the tree, and having the *same* current  $\mathbf{v}_0$  can have values which have size of order  $O(n!^a)$  with some  $a > 0$ . Collecting terms is therefore essential.

A line  $\lambda$  of a tree is said to have scale  $k$  if  $2^{-k-1} \leq 1/C|\omega_0 \cdot \mathbf{v}| < 2^{-k}$ . The lines of a tree  $\theta \in T_n^*$  can then be collected in *clusters*.<sup>1</sup>

A cluster of scale  $p$  is a maximal connected set of lines of scale  $k \geq p$  with at least one line of scale  $p$ . Clusters are connected to the rest of the tree by lines of lower scale which can be *incoming* or *outgoing* with respect to the partial ordering. Clusters also contain nodes: a node is in a cluster if it is an extreme of a line contained in a cluster; such nodes are said to be *internal* to the cluster.

Of particular interest are the *selfenergy* clusters. These are clusters with only one incoming line and only one outgoing line which *furthermore* have the same current  $\mathbf{v}_0$ .

To simplify the analysis the Diophantine condition can be strengthened to insure that if in a tree graph the line incoming into a self energy cluster and ending in an internal node  $v$  is detached from the node  $v$  and reattached to another node internal to the same cluster which is not in a self-energy subcluster (if any) then the new tree nodes are still enclosed in the same clusters. Alternatively the definition of scale of a line can be modified slightly to achieve the same goal.

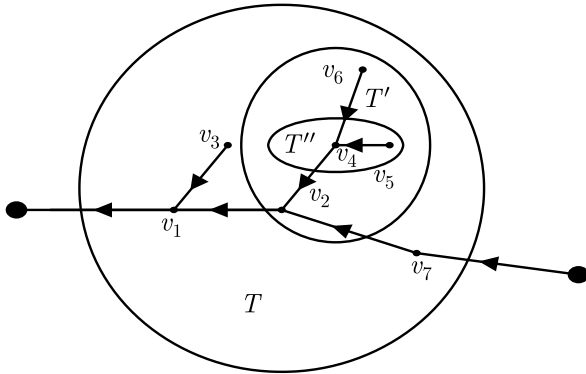
<sup>1</sup>The scaling factor 2 is arbitrary: any scale factor  $> 1$  could be used.

- (3) Then it makes sense to sum together all the values of the trees whose nodes are collected into the same families of clusters and differ only because the lines entering the self energy clusters are attached to a different node internal to the cluster, but external to the inner self energy subclusters (if any). Furthermore, the value of the trees obtained by changing simultaneously sign to the  $v_v$  of the nodes inside the self energy clusters have also to be added together.

After collecting the terms in the described way it is possible to check that each sum of terms so collected is bounded by  $\varrho_0^{-n}$  for some  $\varrho_0$  (which can also be estimated explicitly). Since the number of addends left is not larger than the original one the bound on  $\sum_v |a_{n,v}|$  becomes  $\leq (F^n(2N+1)^\ell n^n N^{2n})/n! \varrho_0^{-n} \leq \varrho^{-n}$ , for suitable  $\varrho_0, \varrho$ , so that convergence of the formal series for  $a_\varepsilon(\varphi)$  is achieved for  $|\varepsilon| < \varrho$ , see [8].

### Non Convergent Cases

Convergence is *not* the rule: very interesting problems arise in which the PT series is, or is believed to be, only asymptotic. For instance in quantum field theory the PT series are well defined but they are not convergent: they can be proved, in the scalar  $\varphi^4$  theories in dimension 2 and 3 to be asymptotic series for a function of  $\varepsilon$  which is *Borel*



**Perturbation Theory, Figure 2**

An example of three clusters symbolically delimited by circles, as visual aids, inside a tree (whose remaining branches and clusters are not drawn and are indicated by the bullets); not all labels are explicitly shown. The scales (not marked) of the branches increase as one crosses inward the circles boundaries: recall, however, that the scale labels are integers  $\leq 1$  (hence typically  $\leq 0$ ). The  $v$  labels are not drawn (but must be imagined). If the  $v$  labels of  $(v_4, v_5)$  add up to 0 the cluster  $T''$  is a self-energy graph. If the  $v$  labels of  $(v_2, v_4, v_5, v_6)$  add up to 0 the cluster  $T'$  is a self-energy graph and such is  $T$  if the  $v$  labels of  $(v_1, v_2, v_3, v_4, v_5, v_6, v_7)$  add up to 0. The cluster  $T'$  is maximal in  $T$ .

*summable*: this means in particular that the solution can be in principle recovered, for  $\varepsilon > 0$  and small, just from the coefficients of its formal expansion.

Other non convergent expansions occur in statistical mechanics, for example in the theory of the ground state of a Fermi gas of particles on a lattice of obstacles. This is still an open problem, and a rather important one. Or occur in quantum field theory where sometimes they can be proved to be Borel summable.

The simplest instances again arise in Mechanics in studying *resonant quasi periodic motions*. A paradigmatic case is provided by Eqs. (1),(2) when  $\omega_0$  has some vanishing components:  $\omega_0 = (\omega_1, \dots, \omega_r, 0, \dots, 0) = (\tilde{\omega}_0, \mathbf{0})$  with  $1 < r < \ell$ . If one writes  $\alpha = (\tilde{\alpha}, \tilde{\beta}) \in \mathcal{T}^r \times \mathcal{T}^{\ell-r}$  and looks motions like Eq. (1) of the form

$$\begin{aligned}\tilde{\alpha}(t) &= \tilde{\varphi} + \tilde{\omega}_0 t + \tilde{a}_\varepsilon(\tilde{\varphi} + \tilde{\omega}_0 t) \\ \tilde{\beta}(t) &= \beta_0 + \tilde{b}_\varepsilon(\tilde{\varphi} + \tilde{\omega}_0 t)\end{aligned}\quad (8)$$

where  $\tilde{a}_\varepsilon(\tilde{\varphi}), \tilde{b}_\varepsilon(\tilde{\varphi})$  are functions of  $\tilde{\varphi} \in \mathcal{T}^r$ , analytic in  $\varepsilon$  and  $\tilde{\varphi}$ .

In this case the analogue of the Lindstedt series can be devised provided  $\beta_0$  is chosen to be a stationary point for the function  $\tilde{f}(\tilde{\beta}) = \int f(\tilde{\alpha}, \tilde{\beta}) \frac{d\tilde{\alpha}}{(2\pi)^r}$ , and provided  $\tilde{\omega}_0$  satisfies a Diophantine property  $|\tilde{\omega}_0 \cdot \tilde{v}| > 1/(C|\tilde{v}|^\tau)$  for all  $\mathbf{0} \neq \tilde{v} \in \mathbb{Z}^r$  and for  $\tau, C$  suitably chosen.

This time the series is likely to be, in general, non convergent (although there is not a proof yet). And the terms of the Lindstedt series can be suitably collected to improve the estimates. Nevertheless, the estimates cannot be improved enough to obtain convergence. Deeper resummations are needed to show that in some cases the terms of the series can be collected and rearranged into a convergent series.

The resummation is deeper in the sense that it is not enough to collect terms contributing to a given order in  $\varepsilon$  but it is necessary to collect and sum terms of different order according to the following scheme.

- (1) The terms of the Lindstedt series are first “regularized” so that the new series is manifestly analytic in  $\varepsilon$  with, however, a radius of convergence depending on the regularization. For instance one can consider only terms with lines of scale  $\leq M$ .
- (2) Terms of different orders in  $\varepsilon$  are then summed together and the series becomes a series in powers of functions  $\lambda_j(\varepsilon; M)$  of  $\varepsilon$  with very small radius of convergence in  $\varepsilon$ , but with an  $M$ -independent radius of convergence  $\varrho$  in the  $\lambda_j(\varepsilon, M)$ . The labels  $j = 0, 1, \dots, M$  are scale labels whose value is determined by the order in which they are generated in

the hierarchical organization of the collection of the graphs according to their scales.

- (3) One shows that the functions  $\lambda_j(\varepsilon; M)$  (“running couplings”) can be analytically continued in  $\varepsilon$  to an  $M$ -independent domain  $\mathcal{D}$  containing the origin in its closure and where they remain smaller than  $\varrho$  for all  $M$ . Furthermore,  $\lambda_j(\varepsilon; M) \xrightarrow{M \rightarrow \infty} \lambda_j(\varepsilon)$ , for  $\varepsilon \in \mathcal{D}$ .
- (4) The convergent power series in the running couplings admits an asymptotic series in  $\varepsilon$  at the origin which coincides with the formal Lindstedt series. Hence in the domain  $\mathcal{D}$  a meaning is attributed to the sum of Lindstedt series.
- (5) One checks that the functions  $\tilde{\mathbf{a}}_\varepsilon, \tilde{\mathbf{b}}_\varepsilon$  thus defined are such that Eq. (8) satisfies the equations of motion Eq. (2).

The proof can be completed if the domain  $\mathcal{D}$  contains real points  $\varepsilon$ .

If  $\tilde{\beta}_0$  is a maximum point the domain  $\mathcal{D}$  contains a circle tangent to the origin and centered on the positive real axis. So in this case the  $\tilde{\mathbf{a}}_\varepsilon, \tilde{\mathbf{b}}_\varepsilon$  are constructed in  $\mathcal{D} \cap \mathcal{R}_+$ ,  $\mathcal{R}_+ \stackrel{\text{def}}{=} (0, +\infty)$ .

If instead  $\tilde{\beta}_0$  is a minimum point the domain  $\mathcal{D}$  exists but  $\mathcal{D} \cap \mathcal{R}_+$  touches the positive real axis on a set of points with positive measure and density 1 at the origin. So  $\tilde{\mathbf{a}}_\varepsilon, \tilde{\mathbf{b}}_\varepsilon$  are constructed only for  $\varepsilon$  in this set which is a kind of “Cantor set” [13].

Again the multiscale analysis is necessary to identify the tree values which have to be collected to define  $\lambda_j(\varepsilon; M)$ . In this case it is an analysis which is much closer to the similar analysis that is encountered in quantum field theory in the “self energy resummations”, which involve collecting and summing graph values of graphs contributing to different orders of perturbation.

The above scheme can also be applied when  $r = \ell$ , i. e. in the case of the classical Lindstedt series when it is actually convergent: this leads to an alternative proof of the Kolmogorov theorem which is interesting as it is even closer to the renormalization group methods because it expresses the solution in terms of a power series in running couplings [Chaps. 8, 9 in 12].

## Conclusion and Outlook

Perturbation theory provides a general approach to the solution of problems “close” to well understood ones, “closeness” being measured by the size of a parameter  $\varepsilon$ . It naturally consists of two steps: the first is to find a formal solution, under the assumption that the quantities of interest are analytic in  $\varepsilon$  at  $\varepsilon = 0$ . If this results in a power series with well-defined coefficients then it becomes necessary to

find whether the series thus constructed, called a *formal series*, converges.

In general the proof that the formal series exists (when it really does) is nontrivial: typically in quantum mechanics problems (quantum fields or statistical mechanics) this is an interesting and deep problem giving rise to renormalization theory. Even in classical mechanics PT of integrable systems it has been, historically, a problem to obtain (in wide generality) the Lindstedt series (of which a simple example is discussed above).

Once existence of a PT series is established, very often the series is not convergent and at best is an asymptotic series. It becomes challenging to find its meaning (if any, as there are cases, even interesting ones, on which conjectures exist claiming that the series have no meaning, like the quantum scalar field in dimension 4 with “ $\varphi^4$ -interaction” or quantum electrodynamics).

Convergence proofs, in most interesting cases, require a multiscale analysis: because the difficulty arises as a consequence of the behavior of singularities at infinitely many scales, as in the case of the Lindstedt series above exemplified.

When convergence is not possible to prove, the multiscale analysis often suggest “resummations”, collecting the various terms whose sums yields the formal PT series (usually the algorithms generating the PT series give its terms at given order as sums of simple but many quantities, as in the discussed case of the Lindstedt series). The collection involves adding together terms of different order in  $\varepsilon$  and results in a new power series, the *resummed series*, in a family of parameters  $\lambda_j(\varepsilon)$  which are functions of  $\varepsilon$ , called the “running couplings”, depending on a “scale index”  $j = 0, 1, \dots$

The running couplings are (in general) singular at  $\varepsilon = 0$  as functions of  $\varepsilon$  but  $C^\infty$  there, and obey equations that allow one to study and define them independently of a convergence proof. If the running couplings can be shown to be so small, as  $\varepsilon$  varies in a suitable domain  $\mathcal{D}$  near 0, to guarantee convergence of the resummed series and therefore to give a meaning to the PT for  $\varepsilon \in \mathcal{D}$  then the PT program can be completed.

The singularities in  $\varepsilon$  at  $\varepsilon = 0$  are therefore all contained in the running couplings, usually very few and the same for various formal series of interest in a given problem.

The idea of expressing the sum of formal series as sum of convergent series in new parameters, the running couplings, determined by other means (a recursion relation denominated the beta function flow) is the key idea of the renormalization group methods: PT in mechanics is a typical and simple example.



On purpose attention has been devoted to PT in the analytic class: but it is possible to use PT techniques in problems in which the functions whose value is studied are not analytic; the techniques are somewhat different and new ideas are needed which would lead quite far away from the natural PT framework which is within the analytic class.

Of course there are many problems of PT in which the formal series are simply convergent and the proof does not require any multiscale analysis. Greater attention was devoted to the novel aspect of PT that emerged in Physics and Mathematics in the last half century and therefore problems not requiring multiscale analysis were not considered. It is worth mentioning, however, that even in simple convergent PT cases it might be convenient to perform resummations. An example is Kepler's equation

$$\ell = \xi - \varepsilon \sin \xi, \quad \xi, \ell \in T^1 = [0, 2\pi] \quad (9)$$

which can be (easily) solved by PT. The resulting series has a radius of convergence in  $\varepsilon$  rather small (Laplace's limit): however if a resummation of the series is performed transforming it into a power series in a "running coupling"  $\lambda_0(\varepsilon)$  (only 1, because no multiscale analysis is needed, the PT series being convergent) given by [Vol. 2, p. 321 in 20]

$$\lambda_0 \stackrel{\text{def}}{=} \frac{\varepsilon e^{\sqrt{1-\varepsilon^2}}}{1 + \sqrt{1-\varepsilon^2}}, \quad (10)$$

then the resummed series is a power series in  $\lambda_0$  with radius of convergence 1 and when  $\varepsilon$  varies between 0 and 1 the parameter  $\lambda_0$  corresponding to it goes from 0 to 1. Hence in terms of  $\lambda_0$  it is possible to invert *by power series* the Kepler equation for all  $\varepsilon \in [0, 1]$ , i. e. in the entire interval of physical interest (recall that  $\varepsilon$  has the interpretation of eccentricity of an elliptic orbit in the 2-body problem). Resummations can improve convergence properties.

### Future Directions

It is always hard to indicate future directions, which usually turn to different paths. Perturbation theory is an ever evolving subject: it is a continuous source of problems and its applications generate new ones. Examples of outstanding problems are understanding the triviality conjectures of models like quantum  $\varphi^4$  field theory in dimension 4 [6]; or a development of the theory of the ground states of Fermionic systems in dimensions 2 and 3 [11]; a theory of weakly coupled Anosov flows to obtain information of the kind that it is possible to obtain for weakly coupled Anosov maps [12]; uniqueness issues in cases in which PT series can be given a meaning, but in a priori non unique

way like the resonant quasi periodic motions in nearly integrable Hamiltonian systems [12].

### Bibliography

1. Einstein A (1917) Zum Quantensatz von Sommerfeld und Epstein. Verh Dtsch Phys Ges 19:82–102
2. Eliasson LH (1996) Absolutely convergent series expansions for quasi-periodic motions. Math Phys Electron J (MPEJ) 2:33
3. Fermi E (1923) Il principio delle adiabatiche e i sistemi che non ammettono coordinate angolari. Nuovo Cimento 25:171–175. Reprinted in Collected papers, vol I, pp 88–91
4. Fröhlich J (1982) On the triviality of  $\lambda\varphi_d^4$  theories and the approach to the critical point in  $d(\geq 4)$  dimensions. Nucl Phys B 200:281–296
5. Gallavotti G (1985) Perturbation Theory for Classical Hamiltonian Systems. In: Fröhlich J (ed) Scaling and self similarity in Physics. Birkhauser, Boston
6. Gallavotti G (1985) Renormalization theory and ultraviolet stability for scalar fields via renormalization group methods. Rev Mod Phys 57:471–562
7. Gallavotti G (1986) Quasi integrable mechanical systems. In: Phénomènes Critiques, Systèmes aleatoires, Théories de jauge (Proceedings, Les Houches, XLIII (1984) vol II, pp 539–624) North Holland, Amsterdam
8. Gallavotti G (1994) Twistless KAM tori. Commun Math Phys 164:145–156
9. Gallavotti G (1995) Invariant tori: a field theoretic point of view on Eliasson's work. In: Figari R (ed) Advances in Dynamical Systems and Quantum Physics. World Scientific, pp 117–132
10. Gallavotti G (2001) Renormalization group in Statistical Mechanics and Mechanics: gauge symmetries and vanishing beta functions. Phys Rep 352:251–272
11. Gallavotti G, Benfatto G (1995) Renormalization group. Princeton University Press, Princeton
12. Gallavotti G, Bonetto F, Gentile G (2004) Aspects of the ergodic, qualitative and statistical theory of motion. Springer, Berlin
13. Gallavotti G, Gentile G (2005) Degenerate elliptic resonances. Commun Math Phys 257:319–362. doi:10.1007/s00220-005-1325-6
14. Glimm J, Jaffe A (1981) Quantum Physics. A functional integral point of view. Springer, New York
15. Gross DJ (1999) Twenty Five Years of Asymptotic Freedom. Nucl Phys B (Proceedings Supplements) 74:426–446. doi:10.1016/S0920-5632(99)00208-X
16. Hepp K (1966) Proof of the Bogoliubov–Parasiuk theorem on renormalization. Commun Math Phys 2:301–326
17. Hepp K (1969) Théorie de la rénormalization. Lecture notes in Physics, vol 2. Springer, Berlin
18. Kolmogorov AN (1954) On the preservation of conditionally periodic motions. Dokl Akad Nauk SSSR 96:527–530 and in: Casati G, Ford J (eds) (1979) Stochastic behavior in classical and quantum Hamiltonians. Lecture Notes in Physics vol 93. Springer, Berlin
19. Laplace PS (1799) Mécanique Céleste. Paris. Reprinted by Chelsea, New York
20. Levi-Civita T (1956) Opere Matematiche, vol 2. Accademia Nazionale dei Lincei and Zanichelli, Bologna
21. Morrey CB (1955) On the derivation of the equations of hydrodynamics from Statistical Mechanics. Commun Pure Appl Math 8:279–326

22. Nelson E (1966) A quartic interaction in two dimensions. In: Goodman R, Segal I (eds) *Mathematical Theory of elementary particles*. MIT, Cambridge, pp 69–73
23. Poincaré H (1892) *Les Méthodes nouvelles de la Mécanique céleste*. Paris. Reprinted by Blanchard, Paris, 1987
24. Pöschel J (1986) Invariant manifolds of complex analytic mappings. In: Osterwalder K, Stora R (eds) *Phénomènes Critiques, Systèmes aleatoires, Théories de jauge* (Proceedings, Les Houches, XLIII (1984); vol II, pp 949–964) North Holland, Amsterdam
25. Siegel K (1943) Iterations of analytic functions. *Ann Math* 43:607–612
26. Simon B (1974) *The  $P(\varphi)_2$  Euclidean (quantum) field theory*. Princeton University Press, Princeton
27. 't Hooft G (1999) When was asymptotic freedom discovered? or The rehabilitation of quantum field theory. *Nucl Phys B (Proceedings Supplements)* 74:413–425. doi:10.1016/S0920-5632(99)00207-8
28. 't Hooft G, Veltman MJG (1972) Regularization and renormalization of gauge fields, *Nucl Phys B* 44:189–213
29. Wilson K, Kogut J (1973) The renormalization group and the  $\varepsilon$ -expansion. *Phys Rep* 12:75–199

## Perturbation Theory in Celestial Mechanics

ALESSANDRA CELLETTI

Dipartimento di Matematica, Università di Roma Tor Vergata, Roma, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Classical Perturbation Theory](#)

[Resonant Perturbation Theory](#)

[Invariant Tori](#)

[Periodic Orbits](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**KAM theory** Provides the persistence of quasi-periodic motions under a small perturbation of an integrable system. KAM theory can be applied under quite general assumptions, i.e. a non-degeneracy of the integrable system and a diophantine condition of the frequency of motion. It yields a constructive algorithm to evaluate the strength of the perturbation ensuring the existence of invariant tori.

**Perturbation theory** Provides an approximate solution of the equations of motion of a nearly-integrable system.

**Spin-orbit problem** A model composed of a rigid satellite rotating about an internal axis and orbiting around a central point-mass planet; a spin-orbit resonance means that the ratio between the revolutional and rotational periods is rational.

**Three-body problem** A system composed by three celestial bodies (e.g. Sun-planet-satellite) assumed to be point-masses subject to the mutual gravitational attraction. The restricted three-body problem assumes that the mass of one of the bodies is so small that it can be neglected.

### Definition of the Subject

Perturbation theory aims to find an approximate solution of nearly-integrable systems, namely systems which are composed by an integrable part and by a small perturbation. The key point of perturbation theory is the construction of a suitable canonical transformation which removes the perturbation to higher orders. A typical example of a nearly-integrable system is provided by a two-body model perturbed by the gravitational influence of a third body whose mass is much smaller than the mass of the central body. Indeed, the solution of the three-body problem greatly stimulated the development of perturbation theories. The solar system dynamics has always been a testing ground for such theories, whose applications range from the computation of the ephemerides of natural bodies to the development of the trajectories of artificial satellites.

### Introduction

The two-body problem can be solved by means of Kepler's laws, according to which for negative energies the point-mass planets move on ellipses with the Sun located in one of the two foci. The dynamics becomes extremely complicated when adding the gravitational influence of another body. Indeed Poincaré showed [34] that the three-body problem does not admit a sufficient number of prime integrals which allow to integrate the problem. Nevertheless, the so-called *restricted* three-body problem deserves special attention, namely when the mass of one of the three bodies is so small that its influence on the others can be neglected. In this case one can assume that the primaries move on Keplerian ellipses around their common barycenter; if the mass of one of the primaries is much larger than the other (as it is the case in any Sun-planet sample), the motion of the minor body is governed by nearly-integrable equations, where the integrable part

represents the interaction with the major body, while the perturbation is due to the influence of the other primary. A typical example is provided by the motion of an asteroid under the gravitational attraction of the Sun and Jupiter. The small body may be taken not to influence the motion of the primaries, which are assumed to move on elliptic trajectories. The dynamics of the asteroid is essentially driven by the Sun and perturbed by Jupiter, since the Jupiter-Sun mass-ratio amounts to about  $10^{-3}$ . The solution of this kind of problem stimulated the work of many scientists, especially in the XVIII and XIX centuries. Indeed, Lagrange, Laplace, Leverrier, Delaunay, Tisserand and Poincaré developed perturbation theories which are the basis of the studies of the dynamics of celestial bodies, from the computation of the ephemerides to the recent advances in flight dynamics. For example, on the basis of perturbation theory Delaunay [16] developed a theory of the Moon, providing very refined ephemerides. Celestial Mechanics greatly motivated the advances of perturbation theories as witnessed by the discovery of Neptune: its position was theoretically predicted by John Adams and by Jean Urbain Leverrier on the basis of perturbative computations; following the suggestion provided by the theoretical investigations, Neptune was finally discovered on 23 September 1846 by the astronomer Johann Gottfried Galle.

The aim of perturbation theory is to implement a canonical transformation which allows one to find the solution of a nearly-integrable system within a better degree of approximation (see Sect. “Classical Perturbation Theory” and references [3,6,20,24,32,37,38]). Let us denote the frequency vector of the system by  $\underline{\omega}$  (see “Normal Forms in Perturbation Theory”, “Kolmogorov–Arnol’d–Moser (KAM Theory)”), which we assume to belong to  $\mathbf{R}^n$ , where  $n$  is the number of degrees of freedom of the system. Classical perturbation theory can be implemented provided that the frequency vector satisfies a non-resonant relation, which means that there does not exist a vector  $\underline{m} \in \mathbf{Z}^n$  such that  $\underline{\omega} \cdot \underline{m} \equiv \sum_{j=1}^n \omega_j m_j = 0$ . In case there exists such commensurability condition, a resonant perturbation theory can be developed as outlined in Sect. “Resonant Perturbation Theory”. In general, the three-body problem (and, more extensively, the  $N$ -body problem) is described by a *degenerate* Hamiltonian system, which means that the integrable part (i. e., the Keplerian approximation) depends on a subset of the action variables. In this case a degenerate perturbation theory must be implemented as explained in Subsect. “Degenerate Perturbation Theory”. For all the above perturbation theories (classical, resonant and degenerate) an application to Celestial Mechanics is given: the precession of the perihelion

of Mercury, orbital resonances within a three-body framework, the precession of the equinoxes.

Even if the non-resonance condition is satisfied, the quantity  $\underline{\omega} \cdot \underline{m}$  can become arbitrarily small, giving rise to the so-called *small divisor* problem; indeed, these terms appear in the denominator of the series defining the canonical transformations necessary to implement perturbation theory and therefore they might prevent the convergence of the series. In order to overcome the small divisor problem, a breakthrough came with the work of Kolmogorov [26], and was later extended to different mathematical settings by Arnold [2] and Moser [33]. The overall theory is known as the acronym KAM theory. As far as concrete estimates on the allowed size of the perturbation are concerned, the original versions of the theory gave discouraging results, which were extremely far from the physical measurements of the parameters involved in the proof. Nevertheless the implementation of computer-assisted KAM proofs allowed one to obtain results which are in good agreement with reality. Concrete estimates with applications to Celestial Mechanics are reported in Sect. “Invariant Tori”.

In the framework of nearly-integrable systems, a very important role is provided by periodic orbits, which might be used to approximate the dynamics of quasi-periodic trajectories; for example, a truncation of the continued fraction expansion of an irrational frequency provides a sequence of rational numbers, which are associated to periodic orbits eventually approximating a quasi-periodic torus. A classical computation of periodic orbits using a perturbative approach is provided in Sect. “Periodic Orbits”, where an application to the determination of the libration in longitude of the Moon is reported.

## Classical Perturbation Theory

### The Classical Theory

Consider a nearly-integrable Hamiltonian function of the form

$$H(\underline{I}, \underline{\varphi}) = h(\underline{I}) + \varepsilon f(\underline{I}, \underline{\varphi}), \quad (1)$$

where  $h$  and  $f$  are analytic functions of  $\underline{I} \in V$  ( $V$  is an open set of  $\mathbf{R}^n$ ) and  $\underline{\varphi} \in \mathbf{T}^n$  ( $\mathbf{T}^n$  is the standard  $n$ -dimensional torus), while  $\varepsilon > 0$  is a small parameter which measures the strength of the perturbation. The aim of perturbation theory is to construct a canonical transformation, which allows to remove the perturbation to higher orders in the perturbing parameter. To this end, let us look for a canonical change of variables (i. e., with symplectic Jacobian matrix)  $C: (\underline{I}, \underline{\varphi}) \rightarrow (\underline{I}', \underline{\varphi}')$ , such that the Hamiltonian (1)

takes the form

$$H'(\underline{I}', \underline{\varphi}') = H \circ C(\underline{I}, \underline{\varphi}) \equiv h'(\underline{I}') + \varepsilon^2 f'(\underline{I}', \underline{\varphi}'), \quad (2)$$

where  $h'$  and  $f'$  denote the new unperturbed Hamiltonian and the new perturbing function, respectively. To achieve such a result we need to proceed along the following steps: build a suitable canonical transformation close to the identity, perform a Taylor series expansion in the perturbing parameter, require that the unknown transformation removes the dependence on the angle variables up to second order terms, and expand in a Fourier series in order to get an explicit form of the canonical transformation.

The change of variables is defined by the equations

$$\begin{aligned} \underline{I} &= \underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} \\ \underline{\varphi}' &= \underline{\varphi} + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{I}'} \end{aligned} \quad (3)$$

where  $\Phi(\underline{I}', \underline{\varphi})$  is an unknown generating function, which is determined so that (1) takes the form (2). Decompose the perturbing function as

$$f(\underline{I}, \underline{\varphi}) = f_0(\underline{I}) + \tilde{f}(\underline{I}, \underline{\varphi}),$$

where  $f_0$  is the average over the angle variables and  $\tilde{f}$  is the remainder function defined through  $\tilde{f}(\underline{I}, \underline{\varphi}) \equiv f(\underline{I}, \underline{\varphi}) - f_0(\underline{I})$ . Define the *frequency vector*  $\underline{\omega} = \underline{\omega}(\underline{I})$  as

$$\underline{\omega}(\underline{I}) \equiv \frac{\partial h(\underline{I})}{\partial \underline{I}}.$$

Inserting the transformation (3) in (1) and expanding in a Taylor series around  $\varepsilon = 0$  up to the second order, one gets

$$\begin{aligned} &h\left(\underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}}\right) + \varepsilon f\left(\underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}}, \underline{\varphi}\right) \\ &= h(\underline{I}') + \underline{\omega}(\underline{I}') \cdot \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} + \varepsilon f_0(\underline{I}') \\ &\quad + \varepsilon \tilde{f}(\underline{I}', \underline{\varphi}) + O(\varepsilon^2). \end{aligned}$$

The new Hamiltonian is integrable up to  $O(\varepsilon^2)$  provided that the function  $\Phi$  satisfies:

$$\underline{\omega}(\underline{I}') \cdot \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} + \tilde{f}(\underline{I}', \underline{\varphi}) = 0. \quad (4)$$

In such case the new integrable part becomes

$$h'(\underline{I}') = h(\underline{I}') + \varepsilon f_0(\underline{I}'),$$

which provides a better integrable approximation with respect to (1). The solution of (4) yields the explicit expression of the generating function. In fact, let us expand  $\Phi$  and  $\tilde{f}$  in Fourier series as

$$\begin{aligned} \Phi(\underline{I}', \underline{\varphi}) &= \sum_{\underline{m} \in \mathbb{Z}^n \setminus \{0\}} \hat{\Phi}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}}, \\ \tilde{f}(\underline{I}', \underline{\varphi}) &= \sum_{\underline{m} \in \mathcal{I}} \hat{f}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}}, \end{aligned} \quad (5)$$

where  $\mathcal{I}$  denotes the set of integer vectors corresponding to the non-vanishing Fourier coefficients of  $\tilde{f}$ . Inserting the above expansions in (4) one obtains

$$i \sum_{\underline{m} \in \mathbb{Z}^n \setminus \{0\}} \underline{\omega}(\underline{I}') \cdot \underline{m} \hat{\Phi}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}} = - \sum_{\underline{m} \in \mathcal{I}} \hat{f}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}},$$

which provides

$$\hat{\Phi}_{\underline{m}}(\underline{I}') = - \frac{\hat{f}_{\underline{m}}(\underline{I}')}{i \underline{\omega}(\underline{I}') \cdot \underline{m}}.$$

Casting together the above formule, the generating function is given by

$$\Phi(\underline{I}', \underline{\varphi}) = i \sum_{\underline{m} \in \mathcal{I}} \frac{\hat{f}_{\underline{m}}(\underline{I}')}{\underline{\omega}(\underline{I}') \cdot \underline{m}} e^{i \underline{m} \cdot \underline{\varphi}}. \quad (6)$$

We stress that this algorithm is constructive in the sense that it provides an explicit expression for the generating function and for the transformed Hamiltonian. We remark that (6) is well defined unless there exists an integer vector  $\underline{m} \in \mathcal{I}$  such that

$$\underline{\omega}(\underline{I}') \cdot \underline{m} = 0.$$

On the contrary, if  $\underline{\omega}$  is rationally independent, there are no zero divisors in (6), though these terms can become arbitrarily small with a proper choice of the vector  $\underline{m}$ . This problem is known as the *small divisor problem*, which can prevent the implementation of perturbation theory (see “Normal Forms in Perturbation Theory”, “Kolmogorov–Arnol’d–Moser (KAM Theory)”, “Perturbation Theory”).

### The Precession of the Perihelion of Mercury

As an example of the implementation of classical perturbation theory we consider the computation of the precession of the perihelion in a (restricted, planar, circular) three-body model, taking as a sample the planet Mercury. The computation requires the introduction of Delaunay action-angle variables, the definition of the three-body Hamiltonian, the expansion of the perturbing function and the implementation of classical perturbation theory (see [7,39]).

**Delaunay Action-Angle Variables** We consider two bodies, say  $P_0$  and  $P_1$  with masses, respectively,  $m_0, m_1$ ; let  $M \equiv m_0 + m_1$  and let  $\mu > 0$  be a positive parameter. Let  $r$  be the orbital radius and  $\varphi$  be the longitude of  $P_1$  with respect to  $P_0$ ; let  $(I_r, I_\varphi)$  be the momenta conjugated to  $(r, \varphi)$ . In these coordinates the two-body problem Hamiltonian takes the form

$$H_{2b}(I_r, I_\varphi, r, \varphi) = \frac{1}{2\mu} \left( I_r^2 + \frac{I_\varphi^2}{r^2} \right) - \frac{\mu M}{r}. \quad (7)$$

On the orbital plane we introduce the planar Delaunay action-angle variables  $(\Lambda, \Gamma, \lambda, \gamma)$  as follows [12]. Let  $E$  denote the total mechanical energy; then:

$$I_r = \sqrt{2\mu E + \frac{2\mu^2 M}{r} - \frac{I_\varphi^2}{r^2}}.$$

Since (7) does not depend on  $\varphi$ , setting  $\Gamma = I_\varphi$  and  $\Lambda = \sqrt{-(\mu^3 M^2)/(2E)}$ , we introduce a generating function of the form

$$F(\Lambda, \Gamma, r, \varphi) = \int \sqrt{-\frac{\mu^4 M^2}{\Lambda^2} + \frac{2\mu^2 M}{r} - \frac{\Gamma^2}{r^2}} dr + \Gamma \varphi.$$

From the definition of  $\Lambda$  the new Hamiltonian  $H_{2D}$  becomes

$$H_{2D}(\Lambda, \Gamma, \lambda, \gamma) = -\frac{\mu^3 M^2}{2\Lambda^2},$$

where  $(\Lambda, \Gamma)$  are the Delaunay action variables; by Kepler's laws one finds that  $(\Lambda, \Gamma)$  are related to the semi-major axis  $a$  and to the eccentricity  $e$  of the Keplerian orbit of  $P_1$  around  $P_0$  by the formula:

$$\Lambda = \mu \sqrt{Ma}, \quad \Gamma = \Lambda \sqrt{1 - e^2}.$$

Concerning the conjugated angle variables, we start by introducing the eccentric anomaly  $u$  as follows: build the auxiliary circle of the ellipse, draw the line through  $P_1$  perpendicular to the semi-major axis whose intersection with the auxiliary circle forms at the origin an angle  $u$  with the semi-major axis. By the definition of the generating function, one finds

$$\begin{aligned} \lambda &= \frac{\partial F}{\partial \Lambda} = \int \frac{\mu^4 M^2}{\Lambda^3 \sqrt{-\frac{\mu^4 M^2}{\Lambda^2} + \frac{2\mu^2 M}{r} - \frac{\Gamma^2}{r^2}}} dr \\ &= u - e \sin u, \end{aligned}$$

which defines the mean anomaly  $\lambda$  in terms of the eccentric anomaly  $u$ .

In a similar way, if  $f$  denotes the true anomaly related to the eccentric anomaly by  $\tan f/2 = \sqrt{(1+e)/(1-e)} \tan u/2$ , then one has:

$$\begin{aligned} \gamma &= \frac{\partial F}{\partial \Gamma} = \varphi - \int \frac{\Gamma}{r^2 \sqrt{-\frac{\mu^4 M^2}{\Lambda^2} + \frac{2\mu^2 M}{r} - \frac{\Gamma^2}{r^2}}} dr \\ &= \varphi - f, \end{aligned}$$

which represents the argument of the perihelion of  $P_1$ , i. e. the angle between the perihelion line and a fixed reference line.

### The Restricted, Planar, Circular, Three-Body Problem

Let  $P_0, P_1, P_2$  be three bodies with masses  $m_0, m_1, m_2$ , respectively. We assume that  $m_1$  is much smaller than  $m_0$  and  $m_2$  (restricted problem) and that the motion of  $P_2$  around  $P_0$  is circular. We also assume that the three bodies always move on the same plane. We choose the free parameter  $\mu$  as  $\mu \equiv 1/m_0^{2/3}$ , so that the two-body Hamiltonian becomes  $H_{2D} = -1/(2\Lambda^2)$ , while we introduce the perturbing parameter as  $\varepsilon \equiv m_2/m_0^{2/3}$  [12]. Set the units of measure so that the distance between  $P_0$  and  $P_2$  is one and so that  $m_0 + m_2 = 1$ . Taking into account the interaction of  $P_2$  on  $P_1$ , the Hamiltonian function governing the three-body problem becomes

$$\begin{aligned} H_{3b}(\Lambda, \Gamma, \lambda, \gamma, t) &= -\frac{1}{2\Lambda^2} \\ &+ \varepsilon \left( r_1 \cos(\varphi - t) - \frac{1}{\sqrt{1 + r_1^2 - 2r_1 \cos(\varphi - t)}} \right), \end{aligned}$$

where  $r_1$  is the distance between  $P_0$  and  $P_1$ . The first term of the perturbation comes out from the choice of the reference frame, while the second term is due to the interaction with the external body. Since  $\varphi - t = f + \gamma - t$ , we perform the canonical change of variables

$$\begin{aligned} L &= \Lambda & \ell &= \lambda \\ G &= \Gamma & g &= \gamma - t, \end{aligned}$$

which provides the following two degrees-of-freedom Hamiltonian

$$H_{3D}(L, G, \ell, g) = -\frac{1}{2L^2} - G + \varepsilon R(L, G, \ell, g), \quad (8)$$

where

$$\begin{aligned} R(L, G, \ell, g) &\equiv r_1 \cos(\varphi - t) - \frac{1}{\sqrt{1 + r_1^2 - 2r_1 \cos(\varphi - t)}} \end{aligned} \quad (9)$$



where  $r_1$  and  $\varphi - t$  must be expressed in terms of the Delaunay variables  $(L, G, \ell, g)$ . Notice that when  $\varepsilon = 0$  one obtains the integrable Hamiltonian function  $h(L, G) \equiv -1/(2L^2) - G$  with associated frequency vector  $\underline{\omega} = (\partial h/\partial L, \partial h/\partial G) = (1/L^3, -1)$ .

**Expansion of the Perturbing Function** We expand the perturbing function (9) in terms of the Legendre polynomials  $P_j$  obtaining

$$R = -\frac{1}{r_1} \sum_{j=2}^{\infty} P_j(\cos(\varphi - t)) \frac{1}{r_1^j}.$$

The explicit expressions of the first few Legendre polynomials are:

$$\begin{aligned} P_2(\cos(\varphi - t)) &= \frac{1}{4} + \frac{3}{4} \cos 2(\varphi - t) \\ P_3(\cos(\varphi - t)) &= \frac{3}{8} \cos(\varphi - t) + \frac{5}{8} \cos 3(\varphi - t) \\ P_4(\cos(\varphi - t)) &= \frac{9}{64} + \frac{5}{16} \cos 2(\varphi - t) \\ &\quad + \frac{35}{64} \cos 4(\varphi - t) \\ P_5(\cos(\varphi - t)) &= \frac{15}{64} \cos(\varphi - t) + \frac{35}{128} \cos 3(\varphi - t) \\ &\quad + \frac{63}{128} \cos 5(\varphi - t). \end{aligned}$$

We invert Kepler's equation  $\ell = u - e \sin u$  to the second order in the eccentricity as

$$u = \ell + e \sin \ell + \frac{e^2}{2} \sin(2\ell) + O(e^3),$$

from which one gets

$$\begin{aligned} \varphi - t &= g + \ell + 2e \sin \ell + \frac{5}{4} e^2 \sin 2\ell + O(e^3) \\ r_1 &= a \left( 1 + \frac{1}{2} e^2 - e \cos \ell - \frac{1}{2} e^2 \cos 2\ell \right) + O(e^3). \end{aligned}$$

Then, up to inessential constants the perturbing function can be expanded as

$$\begin{aligned} R &= R_{00}(L, G) + R_{10}(L, G) \cos \ell + R_{11}(L, G) \cos(\ell + g) \\ &\quad + R_{12}(L, G) \cos(\ell + 2g) + R_{22}(L, G) \cos(2\ell + 2g) \\ &\quad + R_{32}(L, G) \cos(3\ell + 2g) + R_{33}(L, G) \cos(3\ell + 3g) \\ &\quad + R_{44}(L, G) \cos(4\ell + 4g) + R_{55}(L, G) \cos(5\ell + 5g) \\ &\quad + \dots, \end{aligned}$$

where the coefficients  $R_{ij}$  are given by the following expressions (recall that  $e = \sqrt{1 - G^2/L^2}$ ):

$$\begin{aligned} R_{00} &= -\frac{L^4}{4} \left( 1 + \frac{9}{16} L^4 + \frac{3}{2} e^2 \right) + \dots, \\ R_{10} &= \frac{L^4 e}{2} \left( 1 + \frac{9}{8} L^4 \right) + \dots \\ R_{11} &= -\frac{3}{8} L^6 \left( 1 + \frac{5}{8} L^4 \right) + \dots, \\ R_{12} &= \frac{L^4 e}{4} (9 + 5L^4) + \dots \\ R_{22} &= -\frac{L^4}{4} \left( 3 + \frac{5}{4} L^4 \right) + \dots, \\ R_{32} &= -\frac{3}{4} L^4 e + \dots \\ R_{33} &= -\frac{5}{8} L^6 \left( 1 + \frac{7}{16} L^4 \right) + \dots, \\ R_{44} &= -\frac{35}{64} L^8 + \dots \\ R_{55} &= -\frac{63}{128} L^{10} + \dots \end{aligned} \tag{11}$$

**Computation of the Precession of the Perihelion** We identify the three bodies  $P_0, P_1, P_2$  with the Sun, Mercury, and Jupiter, respectively. Taking  $\varepsilon$  as perturbing parameter, we implement a first order perturbation theory, which provides a new integrable Hamiltonian function of the form

$$h'(L', G') = -\frac{1}{2L'^2} - G' + \varepsilon R_{00}(L', G').$$

From Hamilton's equations one obtains

$$\dot{g} = \frac{\partial h'(L', G')}{\partial G'} = -1 + \varepsilon \frac{\partial R_{00}(L', G')}{\partial G'};$$

neglecting  $O(e^3)$  in  $R_{00}$  and recalling that  $g = \gamma - t$ , one has

$$\dot{\gamma} = \varepsilon \frac{\partial R_{00}(L', G')}{\partial G'} = \frac{3}{4} \varepsilon L'^2 G'.$$

Notice that to the first order in  $\varepsilon$  one has  $L' = L$ ,  $G' = G$ . The astronomical data are  $m_0 = 2 \cdot 10^{30}$  kg,  $m_2 = 1.9 \cdot 10^{27}$  kg, which give  $\varepsilon = 9.49 \cdot 10^{-4}$ ; setting to one the Jupiter-Sun distance one has  $a = 0.0744$ , while  $e = 0.2056$ . Taking into account that the orbital period of Jupiter amounts to about 11.86 years, one obtains

$$\dot{\gamma} = 154.65 \frac{\text{arcsecond}}{\text{century}},$$

which represents the contribution due to Jupiter to the precession of perihelion of Mercury. The value found by

Leverrier on the basis of the data available in the year 1856 was of 152.59 arcsecond/century [14].

### Resonant Perturbation Theory

#### The Resonant Theory

Let us consider an Hamiltonian system with  $n$  degrees of freedom of the form

$$H(\underline{I}, \underline{\varphi}) = h(\underline{I}) + \varepsilon f(\underline{I}, \underline{\varphi})$$

and let  $\omega_j(\underline{I}) = (\partial h(\underline{I})/(\partial I_j))$  ( $j = 1, \dots, n$ ) be the frequencies of the motion, which we assume to satisfy  $\ell$ ,  $\ell < n$ , resonance relations of the form

$$\underline{\omega} \cdot \underline{m}_k = 0 \quad \text{for } k = 1, \dots, \ell,$$

for suitable rational independent integer vectors  $\underline{m}_1, \dots, \underline{m}_\ell$ . A resonant perturbation theory can be implemented to eliminate the non-resonant terms. More precisely, the aim is to construct a canonical transformation  $C: (\underline{I}, \underline{\varphi}) \rightarrow (\underline{J}', \underline{\vartheta}')$  such that the transformed Hamiltonian takes the form

$$H'(\underline{J}', \underline{\vartheta}') = h'(\underline{J}', \vartheta'_1, \dots, \vartheta'_\ell) + \varepsilon^2 f'(\underline{J}', \underline{\vartheta}'), \quad (12)$$

where  $h'$  depends only on the resonant angles  $\vartheta'_1, \dots, \vartheta'_\ell$ . To this end, let us first introduce the angles  $\underline{\vartheta} \in \mathbf{T}^n$  as

$$\begin{aligned} \vartheta_j &= \underline{m}_j \cdot \underline{\varphi} & j &= 1, \dots, \ell \\ \vartheta_k &= \underline{m}_k \cdot \underline{\varphi} & k &= \ell + 1, \dots, n, \end{aligned}$$

where the first  $\ell$  angle variables are the resonant angles, while the latter  $n - \ell$  angle variables are defined as suitable linear combinations so to make the transformation canonical together with the following change of coordinates on the actions  $\underline{J} \in \mathbf{R}^n$ :

$$\begin{aligned} I_j &= \underline{m}_j \cdot \underline{J} & j &= 1, \dots, \ell \\ I_k &= \underline{m}_k \cdot \underline{J} & k &= \ell + 1, \dots, n. \end{aligned}$$

The aim is to construct a canonical transformation which removes (to higher order) the dependence on the short-period angles  $(\vartheta_{\ell+1}, \dots, \vartheta_n)$ , while the lowest order Hamiltonian will necessarily depend upon the resonant angles. Let us decompose the perturbation as

$$f(\underline{J}, \underline{\vartheta}) = \bar{f}(\underline{J}) + f_r(\underline{J}, \vartheta_1, \dots, \vartheta_\ell) + f_n(\underline{J}, \underline{\vartheta}), \quad (13)$$

where  $\bar{f}$  is the average of the perturbation over the angles,  $f_r$  is the part depending on the resonant angles and  $f_n$

is the non-resonant part. In analogy to the classical perturbation theory, we implement a canonical transformation of the form

$$\begin{aligned} \underline{J} &= \underline{J}' + \varepsilon \frac{\partial \Phi}{\partial \underline{\vartheta}}(\underline{J}', \underline{\vartheta}) \\ \underline{\vartheta}' &= \underline{\vartheta} + \varepsilon \frac{\partial \Phi}{\partial \underline{J}'}(\underline{J}', \underline{\vartheta}), \end{aligned}$$

such that the new Hamiltonian takes the form (12). Taking into account (13) and developing up to the second order in the perturbing parameter, one obtains:

$$\begin{aligned} h\left(\underline{J}' + \varepsilon \frac{\partial \Phi}{\partial \underline{\vartheta}}\right) + \varepsilon f(\underline{J}', \underline{\vartheta}) + O(\varepsilon^2) \\ = h(\underline{J}') + \varepsilon \sum_{k=1}^n \frac{\partial h}{\partial J_k} \frac{\partial \Phi}{\partial \vartheta_k} + \varepsilon \bar{f}(\underline{J}') + \varepsilon f_r(\underline{J}', \vartheta_1, \dots, \vartheta_\ell) \\ + \varepsilon f_n(\underline{J}', \underline{\vartheta}) + O(\varepsilon^2). \end{aligned}$$

Equating same orders of  $\varepsilon$  one gets that

$$h'(\underline{J}', \vartheta'_1, \dots, \vartheta'_\ell) = h(\underline{J}') + \varepsilon \bar{f}(\underline{J}') + \varepsilon f_r(\underline{J}', \vartheta'_1, \dots, \vartheta'_\ell), \quad (14)$$

provided that

$$\sum_{k=1}^n \omega'_k \frac{\partial \Phi}{\partial \vartheta_k} = -f_n(\underline{J}', \underline{\vartheta}), \quad (15)$$

where  $\omega'_k = \omega'_k(\underline{J}') \equiv (\partial h(\underline{J}')/(\partial J'_k))$ . The solution of (15) gives the generating function, which allows one to reduce the Hamiltonian to the required form (12); as a consequence, the conjugated action variables, say  $J_{\ell+1}', \dots, J_n'$ , are constants of the motion up to the second order in  $\varepsilon$ . We conclude by mentioning that using the new frequencies  $\omega'_k$ , the resonant relations take the form  $\omega'_k = 0$  for  $k = 1, \dots, \ell$ .

#### Three-Body Resonance

We consider the three-body Hamiltonian (8) with perturbing function (10)–(11) and let  $\underline{\omega} \equiv (\omega_\ell, \omega_g)$  be the frequency of motion. We assume that the frequency vector satisfies the resonance relation

$$\omega_\ell + 2\omega_g = 0.$$

According to the theory described in the previous section we perform the canonical change of variables

$$\begin{aligned} \vartheta_1 &= \ell + 2g & J_1 &= \frac{1}{2}G \\ \vartheta_2 &= 2\ell & J_2 &= \frac{1}{2}L - \frac{1}{4}G. \end{aligned}$$

In the new coordinates the unperturbed Hamiltonian becomes

$$h'(\underline{J}) \equiv -\frac{1}{2(J_1 + 2J_2)^2} - 2J_1,$$

with frequency vector  $\underline{\omega}' \equiv \frac{\partial h'(\underline{J})}{\partial \underline{J}}$ , while the perturbation takes the form

$$\begin{aligned} R(J_1, J_2, \vartheta_1, \vartheta_2) &\equiv R_{00}(\underline{J}) + R_{10}(\underline{J}) \cos\left(\frac{1}{2}\vartheta_2\right) \\ &+ R_{11}(\underline{J}) \cos\left(\frac{1}{2}\vartheta_1 + \frac{1}{4}\vartheta_2\right) + R_{12}(\underline{J}) \cos(\vartheta_1) \\ &+ R_{22}(\underline{J}) \cos\left(\vartheta_1 + \frac{1}{2}\vartheta_2\right) + R_{32}(\underline{J}) \cos(\vartheta_1 + \vartheta_2) \\ &+ R_{33}(\underline{J}) \cos\left(\frac{3}{2}\vartheta_1 + \frac{3}{4}\vartheta_2\right) + R_{44}(\underline{J}) \cos(2\vartheta_1 + \vartheta_2) \\ &+ R_{55}(\underline{J}) \cos\left(\frac{5}{2}\vartheta_1 + \frac{5}{4}\vartheta_2\right) + \dots \end{aligned}$$

with the coefficients  $R_{ij}$  as in (11). Let us decompose the perturbation as  $R = \bar{R}(\underline{J}) + R_r(\underline{J}, \vartheta_1) + R_n(\underline{J}, \vartheta)$ , where  $\bar{R}(\underline{J})$  is the average over the angles,  $R_r(\underline{J}, \vartheta_1) = R_{12}(\underline{J}) \cos(\vartheta_1)$  is the resonant part, while  $R_n$  contains all the remaining non-resonant terms. We look for a canonical transformation close to the identity with generating function  $\Phi = \Phi(\underline{J}', \underline{\vartheta})$  such that

$$\underline{\omega}'(\underline{J}') \cdot \frac{\partial \Phi(\underline{J}', \underline{\vartheta})}{\partial \underline{\vartheta}} = -R_n(\underline{J}', \underline{\vartheta}),$$

which is well defined since  $\underline{\omega}'$  is non-resonant for the Fourier components appearing in  $R_n$ . Finally, according to (14) the new unperturbed Hamiltonian is given by

$$h'(\underline{J}', \vartheta'_1) \equiv h(\underline{J}') + \varepsilon R_{00}(\underline{J}') + \varepsilon R_{12}(\underline{J}') \cos \vartheta'_1.$$

### Degenerate Perturbation Theory

A special case of resonant perturbation theory is obtained when considering a degenerate Hamiltonian function with  $n$  degrees of freedom of the form

$$H(\underline{I}, \underline{\varphi}) = h(I_1, \dots, I_d) + \varepsilon f(\underline{I}, \underline{\varphi}), \quad d < n; \quad (16)$$

notice that the integrable part depends on a subset of the action variables, being degenerate in  $I_{d+1}, \dots, I_n$ . In this case we look for a canonical transformation  $C: (\underline{I}, \underline{\varphi}) \rightarrow (\underline{I}', \underline{\varphi}')$  such that the transformed Hamiltonian becomes

$$H'(\underline{I}', \underline{\varphi}') = h'(\underline{I}') + \varepsilon h'_1(\underline{I}, \varphi'_{d+1}, \dots, \varphi'_n) + \varepsilon^2 f'(\underline{I}', \underline{\varphi}'), \quad (17)$$

where the part  $h' + \varepsilon h'_1$  admits  $d$  integrals of motion. Let us decompose the perturbing function in (16) as

$$f(\underline{I}, \underline{\varphi}) = \bar{f}(\underline{I}) + f_d(\underline{I}, \varphi_{d+1}, \dots, \varphi_n) + \tilde{f}(\underline{I}, \underline{\varphi}), \quad (18)$$

where  $\bar{f}$  is the average over the angle variables,  $f_d$  is independent on  $\varphi_1, \dots, \varphi_d$  and  $\tilde{f}$  is the remainder. As in the previous sections we want to determine a near-to-identity canonical transformation  $\Phi = \Phi(\underline{I}', \underline{\varphi})$  of the form (3), such that in view of (18) the Hamiltonian (16) takes the form (17). One obtains

$$\begin{aligned} h(I'_1, \dots, I'_d) + \varepsilon \sum_{k=1}^d \frac{\partial h}{\partial I_k} \frac{\partial \Phi}{\partial \varphi_k} + \varepsilon \bar{f}(I') \\ + \varepsilon f_d(I', \varphi_{d+1}, \dots, \varphi_n) + \varepsilon \tilde{f}(I', \underline{\varphi}) + O(\varepsilon^2) \\ = h'(I') + \varepsilon h'_1(I', \varphi_{d+1}, \dots, \varphi_n) + O(\varepsilon^2), \end{aligned}$$

where

$$\begin{aligned} h'(I') &= h(I'_1, \dots, I'_d) + \varepsilon \bar{f}(I') \\ h'_1(I', \varphi_{d+1}, \dots, \varphi_n) &= f_d(I', \varphi_{d+1}, \dots, \varphi_n), \end{aligned}$$

while  $\Phi$  is determined solving the equation

$$\sum_{k=1}^d \frac{\partial h}{\partial I_k} \frac{\partial \Phi}{\partial \varphi_k} + \tilde{f}(I', \underline{\varphi}) = 0.$$

Expanding  $\Phi$  and  $\tilde{f}$  in Fourier series as in (5) one obtains that  $\Phi$  is given by (6) where  $\underline{\omega} \cdot \underline{m} = \sum_{k=1}^d m_k \omega_k$ , being  $\omega_k = 0$  for  $k = d+1, \dots, n$ . The generating function is well defined provided that  $\underline{\omega} \cdot \underline{m} \neq 0$  for any  $\underline{m} \in \mathcal{I}$ , which is equivalent to requiring that

$$\sum_{k=1}^d m_k \omega_k \neq 0 \quad \text{for } \underline{m} \in \mathcal{I}.$$

### The Precession of the Equinoxes

An example of the application of the degenerate perturbation theory in Celestial Mechanics is provided by the computation of the precession of the equinoxes.

We consider a triaxial rigid body moving in the gravitational field of a primary body. We introduce the following reference frames with a common origin in the barycenter of the rigid body:  $(O, \underline{i}_1^{(i)}, \underline{i}_2^{(i)}, \underline{i}_3^{(i)})$  is an inertial reference frame,  $(O, \underline{i}_1^{(b)}, \underline{i}_2^{(b)}, \underline{i}_3^{(b)})$  is a body frame oriented along the direction of the principal axes of the ellipsoid,  $(O, \underline{i}_1^{(s)}, \underline{i}_2^{(s)}, \underline{i}_3^{(s)})$  is the spin reference frame with the vertical axis along the direction of the angular momentum. Let  $(J, g, \ell)$  be the Euler angles formed by the body and spin

frames, and let  $(K, h, 0)$  be the Euler angles formed by the spin and inertial frames. The angle  $K$  is the obliquity (representing the angle between the spin and inertial vertical axes), while  $J$  is the non-principal rotation angle (representing the angle between the spin and body vertical axes).

This problem is conveniently described in terms of the following set of action-angle variables introduced by Andoyer in [1] (see also [17]). Let  $\underline{M}_0$  be the angular momentum and let  $M_0 \equiv |\underline{M}_0|$ ; the action variables are defined as

$$\begin{aligned} G &\equiv \underline{M}_0 \cdot \dot{\underline{i}}_3^{(s)} = M_0 \\ L &\equiv \underline{M}_0 \cdot \dot{\underline{i}}_3^{(b)} = G \cos J \\ H &\equiv \underline{M}_0 \cdot \dot{\underline{i}}_3^{(i)} = G \cos K, \end{aligned}$$

while the corresponding angle variables are the quantities  $(g, \ell, h)$  introduced before.

We limit ourselves to consider the gyroscopic case in which  $I_1 = I_2 < I_3$  are the principal moments of inertia of the rigid body  $E$  around the primary  $S$ ; let  $m_E$  and  $m_S$  be their masses and let  $|E|$  be the volume of  $E$ . We assume that  $E$  orbits on a Keplerian ellipse around  $S$  with semi-major axis  $a$  and eccentricity  $e$ , while  $\lambda_E$  and  $r_E$  denote the longitude and instantaneous orbital radius (due to the assumption of Keplerian motion  $\lambda_E$  and  $r_E$  are known functions of the time). The Hamiltonian describing the motion of  $E$  around  $S$  is given by [15]

$$\begin{aligned} \mathcal{H}(L, G, H, \ell, g, h, t) \\ = \frac{G^2}{2I_1} + \frac{I_1 - I_3}{2I_1 I_3} L^2 + V(L, G, H, \ell, g, h, t), \end{aligned}$$

where the perturbation is implicitly defined by

$$V \equiv - \int_E \frac{\tilde{G} m_S m_E}{|\underline{r}_E + \underline{x}| |\underline{E}|} d\underline{x},$$

$\tilde{G}$  being the gravitational constant. Setting  $r_E = |\underline{r}_E|$  and  $x = |\underline{x}|$ , we can expand  $V$  using the Legendre polynomials as

$$\begin{aligned} V = - \frac{\tilde{G} m_S m_E}{r_E} \int_E \frac{d\underline{x}}{|\underline{E}|} \left[ 1 - \frac{\underline{x} \cdot \underline{r}_E}{r_E^2} + \frac{1}{2r_E^2} \right. \\ \left. \cdot \left( 3 \frac{(\underline{x} \cdot \underline{r}_E)^2}{r_E^2} - x^2 \right) \right] + O\left(\left(\frac{x}{r_E}\right)^3\right). \end{aligned}$$

We further assume that  $J = 0$  (i. e.  $G = L$ ) so that  $E$  rotates around a principal axis. Let  $G_0$  and  $H_0$  be the initial values of  $G$  and  $H$ ; if  $\alpha$  denotes the angle between  $\underline{r}_E$  and  $\dot{\underline{i}}_3^{(b)}$ , the perturbing function can be written as

$$V = \frac{3}{2} \eta \omega \frac{G_0^2}{H_0} \frac{(1 - e \cos \lambda_E)^3}{(1 - e^2)^3} \cos^2 \alpha$$

with  $\eta = (I_3 - I_1)/I_3$  and  $\omega = (\tilde{G} m_S)/a^3 I_3 H_0/G_0^2$ . Elementary computations show that

$$\cos \alpha = \sin(\lambda_E - h) \sqrt{1 - \frac{H^2}{G^2}}.$$

Neglecting first order terms in the eccentricity, we approximate  $(1 - e \cos \lambda_E)^3/(1 - e^2)^3$  with one. A first order degenerate perturbation theory provides that the new unperturbed Hamiltonian is given by

$$\mathcal{K}(G, H) = \frac{G^2}{2I_3} + \frac{3}{2} \eta \omega \frac{G_0^2}{H_0} \frac{G^2 - H^2}{2G^2}.$$

Therefore the average angular velocity of precession is given by

$$\dot{h} = \frac{\partial \mathcal{K}(G, H)}{\partial H} = -\frac{3}{2} \eta \omega \frac{G_0^2}{H_0} \frac{H}{G^2}.$$

At  $t = 0$  it is

$$\dot{h} = -\frac{3}{2} \eta \omega = -\frac{3}{2} \eta \omega_y^2 \omega_d^{-1} \cos K, \quad (19)$$

where we used  $\omega = \omega_y^2 \omega_d^{-1} \cos K$  with  $\omega_y$  being the frequency of revolution and  $\omega_d$  the frequency of rotation.

In the case of the Earth, the astronomical measurements show that  $\eta = 1/298.25$ ,  $K = 23.45^\circ$ . The contribution due to the Sun is thus obtained by inserting  $\omega_y = 1$  year,  $\omega_d = 1$  day in (19), which yields  $\dot{h}^{(S)} = -2.51857 \cdot 10^{-12}$  rad/sec, corresponding to a retrograde precessional period of 79 107.9 years. A similar computation shows that the contribution of the Moon amounts to  $\dot{h}^{(L)} = -5.49028 \cdot 10^{-12}$  rad/sec, corresponding to a precessional period of 36 289.3 years. The total amount is obtained as the sum of  $\dot{h}^{(S)}$  and  $\dot{h}^{(L)}$ , providing an overall retrograde precessional period of 24 877.3 years.

## Invariant Tori

### Invariant KAM Surfaces

We consider an  $n$ -dimensional nearly-integrable Hamiltonian function

$$H(\underline{I}, \underline{\varphi}) = h(\underline{I}) + \varepsilon f(\underline{I}, \underline{\varphi}),$$

defined in a  $2n$ -dimensional phase space  $\mathcal{M} \equiv V \times \mathbf{T}^n$ , where  $V$  is an open bounded region of  $\mathbf{R}^n$ . A KAM torus associated to  $H$  is an  $n$ -dimensional invariant surface on which the flow is described parametrically by a coordinate  $\underline{\theta} \in \mathbf{T}^n$  such that the conjugated flow is linear, namely

$\underline{\theta} \in \mathbf{T}^n \rightarrow \underline{\theta} + \underline{\omega}t$  where  $\underline{\omega} \in \mathbf{R}^n$  is a Diophantine vector, i. e. there exist  $\gamma > 0$  and  $\tau > 0$  such that

$$|\underline{\omega} \cdot \underline{m}| \geq \frac{\gamma}{|\underline{m}|^\tau}, \quad \forall \underline{m} \in \mathbf{Z}^n \setminus \{0\}.$$

Kolmogorov's theorem [26] (see also "Kolmogorov–Arnol'd–Moser (KAM Theory)") ensures the persistence of invariant tori with diophantine frequency, provided  $\varepsilon$  is sufficiently small and provided the unperturbed Hamiltonian is non-degenerate, i. e. for a given torus  $\{I_0\} \times \mathbf{T}^n \subset \mathcal{M}$

$$\det h''(I_0) \equiv \det \left( \frac{\partial^2 h}{\partial I_i \partial I_j}(I_0) \right)_{i,j=1,\dots,n} \neq 0. \quad (20)$$

The condition (20) can be replaced by the isoenergetic non-degeneracy condition introduced by Arnold [2]

$$\det \begin{pmatrix} h''(I_0) & h'(I_0) \\ h'(I_0) & \underline{0} \end{pmatrix} \neq 0, \quad (21)$$

which ensures the existence of KAM tori on the energy level corresponding to the unperturbed energy  $h(I_0)$ , say  $\mathcal{M}_0 \equiv \{(\underline{I}, \underline{\varphi}) \in \mathcal{M} : H(\underline{I}, \underline{\varphi}) = h(I_0)\}$ . In the context of the  $n$ -body problem Arnold [2] addressed the question of the existence of a set of initial conditions with positive measure such that, if the initial position and velocities of the bodies belong to this set, then the mutual distances remain perpetually bounded. A positive answer is provided by Kolmogorov's theorem in the framework of the planar, circular, restricted three-body problem, since the integrable part of the Hamiltonian (8) satisfies the isoenergetic non-degeneracy condition (21); denoting the initial values of the Delaunay's action variables by  $(I_0, G_0)$ , if  $\varepsilon$  is sufficiently small, there exist KAM tori for (8) on the energy level  $\mathcal{M}_0 \equiv \{H_{3D} = -1/(2I_0^2) - G_0\}$ . In particular, the motion of the perturbed body remains forever bounded from the orbits of the primaries. Indeed, a stronger statement is also valid: due to the fact that the two-dimensional KAM surfaces separate the three dimensional energy levels, any trajectory starting between two KAM tori remains forever trapped in the region between such tori.

In the framework of the three-body problem, Arnold [2] stated the following result: "If the masses, eccentricities and inclinations of the planets are sufficiently small, then for the majority of initial conditions the true motion is conditionally periodic and differs little from Lagrangian motion with suitable initial conditions throughout an infinite interval time  $-\infty < t < \infty$ ". Arnold provided a complete proof for the case of three coplanar bodies, while the spatial three-body problem was investigated by Laskar and

Robutel in [27,35] using Poincaré variables, the Jacobi's "reduction of the nodes" (see, e. g., [11]) and Birkhoff's normal form [3,4,38]. The full proof of Arnold's theorem was provided in [19], based on Herman's results on the planetary problem; it makes use of Poincaré variables restricted to the symplectic manifold of vertical total angular momentum.

Explicit estimates on the perturbing parameter ensuring the existence of KAM tori were given by M. Hénon [25]; he showed that direct applications of KAM theory to the three-body problem lead to analytical results which are much smaller than the astronomical observations. For example, the application of Arnold's theorem to the restricted three-body problem is valid provided the mass-ratio of the primaries is less than  $10^{-33}$ . This result can be improved up to  $10^{-48}$  by applying Moser's theorem, but it is still very far from the actual Jupiter-Sun mass-ratio which amounts to about  $10^{-3}$ . In the context of concrete estimates, a big improvement comes from the synergy between KAM theory and computer-assisted proofs, based on the application of *interval arithmetic* which allows to keep rigorously track of the rounding-off and propagation errors introduced by the machine. Computer-assisted KAM estimates were implemented in a number of cases in Celestial Mechanics, like the three-body problem and the spin-orbit model as briefly recalled in the following subsections.

Another interesting example of the interaction between the analytical theory and the computer implementation is provided by the analysis of the stability of the triangular Lagrangian points; in particular, the stability for exponentially long times is obtained using Nekhoroshev theory combined with computer-assisted implementations of Birkhoff normal form (see, e. g., [5,13,18,21,22,23,28,36]).

### Rotational Tori for the Spin-Orbit Problem

We study the motion of a rigid triaxial satellite around a central planet under the following assumptions [8]:

- i) The orbit of the satellite is Keplerian,
- ii) The spin-axis is perpendicular to the orbital plane,
- iii) The spin-axis coincides with the smallest physical axis,
- iv) External perturbations as well as dissipative forces are neglected.

Let  $I_1 < I_2 < I_3$  be the principal moments of inertia; let  $a, e$  be the semi-major axis and eccentricity of the Keplerian ellipse; let  $r$  and  $f$  be the instantaneous orbital radius and the true anomaly of the satellite; let  $x$  be the angle between the longest axis of the triaxial satellite and the periaxis line. The equation of motion governing the spin-



orbit model is given by:

$$\ddot{x} + \frac{3}{2} \frac{I_2 - I_1}{I_3} \left( \frac{a}{r} \right)^3 \sin(2x - 2f) = 0. \quad (22)$$

Due to *assumption i*, the quantities  $r$  and  $f$  are known functions of the time. Expanding the second term of (22) in Fourier-Taylor series and neglecting terms of order 6 in the eccentricity, setting  $y \equiv \dot{x}$  one obtains that the equation of motion corresponds to Hamilton's equations associated to the Hamiltonian

$$\begin{aligned} H(y, x, t) \equiv & \frac{y^2}{2} - \varepsilon \left[ \left( -\frac{e}{4} + \frac{e^3}{32} - \frac{5}{768} e^5 \right) \cos(2x - t) \right. \\ & + \left( \frac{1}{2} - \frac{5}{4} e^2 + \frac{13}{32} e^4 \right) \cos(2x - 2t) \\ & + \left( \frac{7}{4} e - \frac{123}{32} e^3 + \frac{489}{256} e^5 \right) \cos(2x - 3t) \\ & + \left( \frac{17}{4} e^2 - \frac{115}{12} e^4 \right) \cos(2x - 4t) \\ & + \left( \frac{845}{96} e^3 - \frac{32525}{1536} e^5 \right) \cos(2x - 5t) \\ & + \frac{533}{32} e^4 \cos(2x - 6t) \\ & \left. + \frac{228347}{7680} e^5 \cos(2x - 7t) \right], \end{aligned} \quad (23)$$

where  $\varepsilon \equiv 3/2(I_2 - I_1)/I_3$  and we have chosen the units so that  $a = 1$ ,  $2\pi/T_{\text{rev}} = 1$ , where  $T_{\text{rev}}$  is the period of revolution. Let  $p, q$  be integers with  $q \neq 0$ ; a  $p : q$  resonance occurs whenever  $\langle \dot{x} \rangle = \frac{p}{q}$ , meaning that during  $q$  orbital revolutions, the satellite makes on average  $p$  rotations. Since the phase-space is three-dimensional, the two-dimensional KAM tori separates the phase-space into invariant regions, thus providing the stability of the trapped orbits. In particular, let  $\mathcal{P}(\frac{p}{q})$  be the periodic orbit associated to the  $p : q$  resonance; its stability is guaranteed by the existence of trapping rotational tori with frequencies  $\mathcal{T}(\omega_1)$  and  $\mathcal{T}(\omega_2)$  with  $\omega_1 < p/q < \omega_2$ . For example, one can consider the sequences of irrational rotation numbers

$$\Gamma_k^{(p/q)} \equiv \frac{p}{q} - \frac{1}{k + \alpha}, \quad \Delta_k^{(p/q)} \equiv \frac{p}{q} + \frac{1}{k + \alpha},$$

$$k \in \mathbb{Z}, k \geq 2$$

with  $\alpha \equiv (\sqrt{5} - 1)/2$ . In fact, the continued fraction expansion of  $1/(k + \alpha)$  is given by  $1/(k + \alpha) = [0, k, 1^\infty]$ . Therefore, both  $\Gamma_k^{(p/q)}$  and  $\Delta_k^{(p/q)}$  are *noble* numbers (i. e. with continued fraction expansion definitely equal to one);

by number theory they satisfy the diophantine condition and bound  $\frac{p}{q}$  from below and above.

As a concrete sample we consider the synchronous spin-orbit resonance ( $p = q = 1$ ) of the Moon, whose physical values of the parameters are  $\varepsilon \equiv 3.45 \cdot 10^{-4}$  and  $e = 0.0549$ . The stability of the motion is guaranteed by the existence of the surfaces  $\mathcal{T}(\Gamma_{40}^{(1)})$  and  $\mathcal{T}(\Delta_{40}^{(1)})$ , which is obtained by implementing a computer-assisted KAM theory for the realistic values of the parameters. The result provides the confinement of the synchronous periodic orbit in a limited region of the phase space [8].

### Librational Tori for the Spin-Orbit Problem

The existence of invariant librational tori around a spin-orbit resonance can be obtained as follows [9]. Let us consider the 1:1 resonance corresponding to Hamilton's equations associated to (23). First one implements a canonical transformation to center around the synchronous periodic orbit; after expanding in Taylor series, one diagonalizes the quadratic terms, thus obtaining a harmonic oscillator plus higher degree (time-dependent) terms. Finally, it is convenient to transform the Hamiltonian using the action-angle variables  $(I, \varphi)$  of the harmonic oscillator. After these symplectic changes of variables one is led to a Hamiltonian of the form

$$H(I, \varphi, t) \equiv \omega I + \varepsilon \bar{h}(I) + \varepsilon R(I, \varphi, t),$$

$$I \in \mathbb{R}, (\varphi, t) \in T^2,$$

where  $\omega \equiv \omega(\varepsilon)$  is the frequency of the harmonic oscillator, while  $\bar{h}(I)$  and  $R(I, \varphi, t)$  are suitable functions, precisely polynomials in the action (of order of the square of the action). Then apply a Birkhoff normal form (see – Normal Forms in Perturbation Theory –) up to the order  $k$  ( $k = 5$  in [9]) to obtain the following Hamiltonian:

$$H_k(I', \varphi', t) = \omega I' + \varepsilon h_k(I'; \varepsilon) + \varepsilon^{k+1} R_k(I', \varphi', t).$$

Finally, implementing a computer-assisted KAM theorem one gets the following result: consider the Moon–Earth case with  $\varepsilon_{\text{obs}} = 3.45 \cdot 10^{-4}$  and  $e = 0.0549$ ; there exists an invariant torus around the synchronous resonance corresponding to a libration of  $8.79^\circ$  for any  $\varepsilon \leq \varepsilon_{\text{obs}}/5.26$ . The same strategy applied to different samples, e.g. the Rhea–Saturn pair, allows one to prove the existence of librational invariant tori around the synchronous resonance for values of the parameters in full agreement with the observational measurements [9].

### Rotational Tori for the Restricted Three-Body Problem

The planar, circular, restricted three-body problem has been considered in [12], where the stability of the asteroid 12 Victoria has been investigated under the gravitational influence of the Sun and Jupiter. On a fixed energy level invariant KAM tori trapping the motion of Victoria have been established for the astronomical value of the Jupiter–Sun mass-ratio (about  $10^{-3}$ ). After an expansion of the perturbing function and a truncation to a suitable order (see [12]), the Hamiltonian function describing the motion of the asteroid is given in Delaunay’s variables by

$$H(L, G, \ell, g) \equiv -\frac{1}{2L^2} - G - \varepsilon f(L, G, \ell, g),$$

where setting  $a \equiv L^2$ ,  $e = \sqrt{1 - G^2/L^2}$ , the perturbation is given by

$$\begin{aligned} f(L, G, \ell, g) = & 1 + \frac{a^2}{4} + \frac{9}{64} a^4 + \frac{3}{8} a^2 e^2 \\ & - \left( \frac{1}{2} + \frac{9}{16} a^2 \right) a^2 e \cos \ell \\ & + \left( \frac{3}{8} a^3 + \frac{15}{64} a^5 \right) \cos(\ell + g) \\ & - \left( \frac{9}{4} + \frac{5}{4} a^2 \right) a^2 e \cos(\ell + 2g) \\ & + \left( \frac{3}{4} a^2 + \frac{5}{16} a^4 \right) \cos(2\ell + 2g) \\ & + \frac{3}{4} a^2 e \cos(3\ell + 2g) \\ & + \left( \frac{5}{8} a^3 + \frac{35}{128} a^5 \right) \cos(3\ell + 3g) \\ & + \frac{35}{64} a^4 \cos(4\ell + 4g) \\ & + \frac{63}{128} a^5 \cos(5\ell + 5g). \end{aligned}$$

For the asteroid Victoria the orbital elements are  $a_V \simeq 2.334$  AU,  $e_V \simeq 0.220$ , which give the observed values of the Delaunay’s action variables as  $L_V = 0.670$ ,  $G_V = 0.654$ . The energy level is taken as

$$\begin{aligned} E_V^{(0)} & \equiv -\frac{1}{2L_V^2} - G_V \simeq -1.768, \\ E_V^{(1)} & \equiv -\langle f(L_V, G_V, \ell, g) \rangle \simeq -1.060, \\ E_V(\varepsilon) & \equiv E_V^{(0)} + \varepsilon E_V^{(1)}. \end{aligned}$$

The osculating energy level of the Sun–Jupiter–Victoria model is defined as

$$E_V^* \equiv E_V(\varepsilon_J) = E_V^{(0)} + \varepsilon_J E_V^{(1)} \simeq -1.769.$$

We now look for two invariant tori bounding the observed values of  $L_V$  and  $G_V$ . To this end, let  $\tilde{L}_\pm = L_V \pm 0.001$  and let

$$\tilde{\omega}_\pm = \left( \frac{1}{\tilde{L}_\pm^3}, -1 \right) \equiv (\tilde{\alpha}_\pm, -1).$$

To obtain diophantine frequencies, the continued fraction expansion of  $\tilde{\alpha}_\pm$  is modified adding a tail of ones after the order 5; this procedure gives the diophantine numbers  $\alpha_\pm$  which define the bounding frequencies as  $\omega_\pm = (\alpha_\pm, -1)$ . By a computer-assisted KAM theorem, the stability of the asteroid Victoria is a consequence of the following result [12]: for  $|\varepsilon| \leq 10^{-3}$  the unperturbed tori can be analytically continued into invariant KAM tori for the perturbed system on the energy level  $H^{-1}(E_V(\varepsilon))$ , keeping fixed the ratio of the frequencies. Therefore the orbital elements corresponding to the semi-major axis and to the eccentricity of the asteroid Victoria stay forever  $\varepsilon$ -close to their unperturbed values.

### Planetary Problem

The dynamics of the planetary problem composed by the Sun, Jupiter and Saturn is investigated in [29,30,31]. In [29] the secular dynamics of the following model is studied: after the Jacobi’s reduction of the nodes, the 4-dimensional Hamiltonian is averaged over the fast angles and its series expansion is considered up to the second order in the masses. This procedure provides a Hamiltonian function with two degrees of freedom, describing the slow motion of the parameters characterizing the Keplerian approximation (i. e., the eccentricities and the arguments of perihelion). Afterwards, action-angle coordinates are introduced and a partial Birkhoff normalization is performed. Finally, a computer-assisted implementation of a KAM theorem yields the existence of two invariant tori bounding the secular motions of Jupiter and Saturn for the observed values of the parameters.

The approach sketched above is extended in [31] so to include the description of the fast variables, like the semi-major axes and the mean longitudes of the planets. Indeed, the preliminary average on the fast angles is now performed without eliminating the terms with degree greater or equal than two with respect to the fast actions. The canonical transformations involving the secular coordinates can be adapted to produce a good initial approximation of an invariant torus for the reduced Hamiltonian of the three-body planetary problem. This is the starting point of the procedure for constructing the Kolmogorov’s normal form which is numerically shown to be convergent. In [30] the same result of [31] has been obtained for

a fictitious planetary solar system composed by two planets with masses equal to 1/10 of those of Jupiter and Saturn.

## Periodic Orbits

### Construction of Periodic Orbits

One of the most intriguing conjectures of Poincaré concerns the pivotal role of the periodic orbits in the study of the dynamics; more precisely, he states that given a particular solution of Hamilton's equations one can always find a periodic solution (possibly with very long period) such that the difference between the two solutions is small for an arbitrary long time. The literature on periodic orbits is extremely wide (see, e.g., [4,7,24,38,39] and references therein); here we present the construction of periodic orbits implementing a perturbative approach (see [10]) as shown by Poincaré in [34]. We describe such a method taking as an example the spin-orbit Hamiltonian (23) that we write in a compact form as  $H(y, x, t) \equiv y^2/2 - \varepsilon f(x, t)$  for a suitable function  $f = f(x, t)$ ; the corresponding Hamilton's equations are

$$\begin{aligned}\dot{x} &= y \\ \dot{y} &= \varepsilon f_x(x, t).\end{aligned}\quad (24)$$

A spin-orbit resonance of order  $p : q$  is a periodic solution of period  $T = 2\pi q$  ( $q \in \mathbb{Z} \setminus \{0\}$ ), such that

$$\begin{aligned}x(t + 2\pi q) &= x(t) + 2\pi p \\ y(t + 2\pi q) &= y(t).\end{aligned}\quad (25)$$

From (24) the solution can be written in integral form as

$$\begin{aligned}y(t) &= y(0) + \varepsilon \int_0^t f_x(x(s), s) ds \\ x(t) &= x(0) + y(0)t + \varepsilon \int_0^t \int_0^\tau f_x(x(s), s) ds d\tau \\ &= x(0) + \int_0^t y(s) ds;\end{aligned}$$

combining the above equations with (25) one obtains

$$\begin{aligned}\int_0^{2\pi q} f_x(x(s), s) ds &= 0 \\ \int_0^{2\pi q} y(s) ds - 2\pi p &= 0.\end{aligned}\quad (26)$$

Let us write the solution as the series

$$\begin{aligned}x(t) &\equiv \bar{x} + \bar{y}t + \varepsilon x_1(t) + \dots \\ y(t) &\equiv \bar{y} + \varepsilon y_1(t) + \dots,\end{aligned}\quad (27)$$

where  $x(0) = \bar{x}$  and  $y(0) = \bar{y}$  are the initial conditions, while  $x_1(t)$ ,  $y_1(t)$  are the first order terms in  $\varepsilon$ . Expanding the initial conditions in power series of  $\varepsilon$ , one gets:

$$\begin{aligned}\bar{x} &= \bar{x}_0 + \varepsilon \bar{x}_1 + \varepsilon^2 \bar{x}_2 + \dots \\ \bar{y} &= \bar{y}_0 + \varepsilon \bar{y}_1 + \varepsilon^2 \bar{y}_2 + \dots\end{aligned}\quad (28)$$

Inserting (27) and (28) in (24), equating same orders in  $\varepsilon$  and taking into account the periodicity condition (26), one can find the following explicit expressions for  $x_1(t)$ ,  $y_1(t)$ ,  $\bar{y}_0$ ,  $\bar{y}_1$ :

$$\begin{aligned}y_1(t) &= y_1(t; \bar{y}, \bar{x}) = \int_0^t f_x(\bar{x}_0 + \bar{y}_0 s, s) ds \\ x_1(t) &= x_1(t; \bar{y}, \bar{x}) = \int_0^t y_1(s) ds \\ \bar{y}_0 &= \frac{p}{q} \\ \bar{y}_1 &= -\frac{1}{2\pi q} \int_0^{2\pi q} \int_0^t f_x(\bar{x}_0 + \bar{y}_0 s, s) ds dt.\end{aligned}$$

Furthermore,  $\bar{x}_0$  is determined as a solution of

$$\int_0^{2\pi q} f_x(\bar{x}_0 + \bar{y}_0 s, s) ds = 0,$$

while  $\bar{x}_1$  is given by

$$\begin{aligned}\bar{x}_1 &= -\frac{1}{\int_0^{2\pi q} f_{xx}^0 dt} \\ &\quad \cdot \left[ \bar{y}_1 \int_0^{2\pi q} t f_{xx}^0 dt + \int_0^{2\pi q} f_{xx}^0 x_1(t) dt \right],\end{aligned}$$

where, for shortness, we have written  $f_{xx}^0 = f_{xx}(\bar{x}_0 + \bar{y}_0 t, t)$ .

### The Libration in Longitude of the Moon

The previous computation of the  $p : q$  periodic solution can be used to evaluate the libration in longitude of the Moon. More precisely, setting  $p = q = 1$  one obtains

$$\begin{aligned}\bar{x}_0 &= 0 \\ \bar{y}_0 &= 1 \\ x_1(t) &= 0.232086t - 0.218318 \sin(t) \\ &\quad - 6.36124 \cdot 10^{-3} \sin(2t) - 3.21314 \cdot 10^{-4} \sin(3t) \\ &\quad - 1.89137 \cdot 10^{-5} \sin(4t) - 1.18628 \cdot 10^{-6} \sin(5t) \\ y_1(t) &= 0.232086 - 0.218318 \cos(t) - 0.0127225 \cos(2t) \\ &\quad - 9.63942 \cdot 10^{-4} \cos(3t) - 7.56548 \cdot 10^{-5} \cos(4t) \\ &\quad - 5.93138 \cdot 10^{-6} \cos(5t) \\ \bar{x}_1 &= 0 \\ \bar{y}_1 &= -0.232086,\end{aligned}$$

where we used  $e = 0.05494$ ,  $\varepsilon = 3.45 \cdot 10^{-4}$ . Therefore the synchronous periodic solution, computed up to the first order in  $\varepsilon$ , is given by

$$\begin{aligned} x(t) &= \bar{x}_0 + \bar{y}_0 t + \varepsilon x_1(t) = t - 7.53196 \cdot 10^{-5} \sin(t) \\ &\quad - 2.19463 \cdot 10^{-6} \sin(2t) - 1.10853 \cdot 10^{-7} \sin(3t) \\ &\quad - 6.52523 \cdot 10^{-9} \sin(4t) - 4.09265 \cdot 10^{-10} \sin(5t) \\ y(t) &= \bar{y}_0 t + \varepsilon y_1(t) = 1 - 7.53196 \cdot 10^{-5} \cos(t) \\ &\quad - 4.38926 \cdot 10^{-6} \cos(2t) - 3.3256 \cdot 10^{-7} \cos(3t) \\ &\quad - 2.61009 \cdot 10^{-8} \cos(4t) - 2.04633 \cdot 10^{-9} \cos(5t). \end{aligned}$$

It turns out that the libration in longitude of the Moon, provided by the quantity  $x(t) - t$ , is of the order of  $7 \cdot 10^{-5}$  in agreement with the observational data.

### Future Directions

The last decade of the XX century has been greatly marked by astronomical discoveries, which changed the shape of the solar system as well as of the entourage of other stars. In particular, the detection of many small bodies beyond the orbit of Neptune has moved the edge of the solar system forward and it has increased the number of its population. Hundreds of objects have been observed to move in a ring beyond Neptune, thus forming the so-called Kuiper's belt. Its components show a great variety of behaviors, like resonance clusterings, regular orbits, scattered trajectories. Furthermore, far outside the solar system, the astronomical observations of extrasolar planetary systems have opened new scenarios with a great variety of dynamical behaviors. In these contexts classical and resonant perturbation theories will deeply contribute to provide a fundamental insight of the dynamics and will play a prominent role in explaining the different configurations observed within the Kuiper's belt as well as within extrasolar planetary systems.

### Bibliography

- Andoyer H (1926) *Mécanique Céleste*. Gauthier-Villars, Paris
- Arnold VI (1963) Small denominators and problems of stability of motion in classical and celestial mechanics. *Uspehi Mat Nauk* 6 18(114):91–192
- Arnold VI (1978) *Mathematical methods of classical mechanics*. Springer, Berlin
- Arnold VI (ed) (1988) *Encyclopedia of Mathematical Sciences. Dynamical Systems III*. Springer, Berlin
- Benettin G, Fasso F, Guzzo M (1998) Nekhoroshev-stability of  $L_4$  and  $L_5$  in the spatial restricted three-body problem. *Regul Chaotic Dyn* 3(3):56–71
- Boccaletti D, Pucacco G (2001) *Theory of orbits*. Springer, Berlin
- Brouwer D, Clemence G (1961) *Methods of Celestial Mechanics*. Academic Press, New York
- Celletti A (1990) Analysis of resonances in the spin-orbit problem. In: *Celestial Mechanics: The synchronous resonance (Part I)*. *J Appl Math Phys (ZAMP)* 41:174–204
- Celletti A (1993) Construction of librational invariant tori in the spin-orbit problem. *J Appl Math Phys (ZAMP)* 45:61–80
- Celletti A, Chierchia L (1998) Construction of stable periodic orbits for the spin-orbit problem of Celestial Mechanics. *Regul Chaotic Dyn (Editorial URSS)* 3:107–121
- Celletti A, Chierchia L (2006) KAM tori for N-body problems: a brief history. *Celest Mech Dyn Astron* 95 1:117–139
- Celletti A, Chierchia L (2007) KAM Stability and Celestial Mechanics. *Mem Am Math Soc* 187:878
- Celletti A, Giorgilli A (1991) On the stability of the Lagrangian points in the spatial restricted problem of three bodies. *Celest Mech Dyn Astron* 50:31–58
- Chebotaev AG (1967) *Analytical and Numerical Methods of Celestial Mechanics*. Elsevier, New York
- Chierchia L, Gallavotti G (1994) Drift and diffusion in phase space. *Ann l'Inst H Poincaré* 60:1–144
- Delaunay C (1867) *Mémoire sur la théorie de la Lune*. *Mém l'Acad Sci* 28:29 (1860)
- Deprit A (1967) Free rotation of a rigid body studied in the phase space. *Am J Phys* 35:424–428
- Efthymiopoulos C, Sandor Z (2005) Optimized Nekhoroshev stability estimates for the Trojan asteroids with a symplectic mapping model of co-orbital motion. *MNRAS* 364(6):253–271
- Féjóz J (2004) Démonstration du “théorème d'Arnold” sur la stabilité du système planétaire (d'après Michael Herman). *Ergod Th Dynam Syst* 24:1–62
- Ferraz-Mello S (2007) *Canonical Perturbation Theories*. Springer, Berlin
- Gabern F, Jorba A, Locatelli U (2005) On the construction of the Kolmogorov normal form for the Trojan asteroids. *Nonlinearity* 18:1705–1734
- Giorgilli A, Skokos C (1997) On the stability of the trojan asteroids. *Astron Astroph* 317:254–261
- Giorgilli A, Delshams A, Fontich E, Galgani L, Simó C (1989) Effective stability for a Hamiltonian system near an elliptic equilibrium point, with an application to the restricted three-body problem. *J Diff Eq* 77:167–198
- Hagihara Y (1970) *Celestial Mechanics*. MIT Press, Cambridge
- Hénon M (1966) Explorations numériques du problème restreint IV: Masses égales, orbites non périodique. *Bull Astron* 3(1, fasc 2):49–66
- Kolmogorov AN (1954) On the conservation of conditionally periodic motions under small perturbation of the Hamiltonian. *Dokl Akad Nauk SSR* 98:527–530
- Laskar J, Robutel P (1995) Stability of the planetary three-body problem I Expansion of the planetary Hamiltonian. *Celest Mech and Dyn Astron* 62(3):193–217
- Lhotka Ch, Efthymiopoulos C, Dvorak R (2008) Nekhoroshev stability at  $L_4$  or  $L_5$  in the elliptic restricted three body problem-application to Trojan asteroids. *MNRAS* 384:1165–1177
- Locatelli U, Giorgilli A (2000) Invariant tori in the secular motions of the three-body planetary systems. *Celest Mech and Dyn Astron* 78:47–74
- Locatelli U, Giorgilli A (2005) Construction of the Kolmogorov's normal form for a planetary system. *Regul Chaotic Dyn* 10:153–171
- Locatelli U, Giorgilli A (2007) Invariant tori in the Sun–Jupiter–Saturn system. *Discret Contin Dyn Syst-Ser B* 7:377–398

32. Meyer KR, Hall GR (1991) Introduction to Hamiltonian dynamical systems and the  $N$ -body problem. Springer, Berlin
33. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nach Akad Wiss Göttingen. Math Phys Kl II 1:1
34. Poincaré H (1892) Les Méthodes Nouvelles de la Mécanique Céleste. Gauthier-Villars, Paris
35. Robutel P (1995) Stability of the planetary three-body problem II KAM theory and existence of quasi-periodic motions. Celest Mech Dyn Astron 62(3):219–261
36. Robutel P, Gabern F (2006) The resonant structure of Jupiter's Trojan asteroids – I Long-term stability and diffusion. MNRAS 372(4):1463–1482
37. Sanders JA, Verhulst F (1985) Averaging methods in nonlinear dynamical systems. Springer, Berlin
38. Siegel CL, Moser JK (1971) Lectures on Celestial Mechanics. Springer, Heidelberg
39. Szebehely V (1967) Theory of orbits. Academic Press, New York

## Perturbation Theory, Introduction to

GIUSEPPE GAETA

Dipartimento di Matematica,  
Università di Milano, Milan, Italy

The idea behind Perturbation Theory is that when we are not able to determine exact solutions to a given problem, we might be able to determine approximate solutions to our problem starting from solutions to an approximate version of the problem, amenable to exact treatment. Thus, in a way, we use *exact solutions to an approximate problem to get approximate solutions to an exact problem*.

It goes without saying that many mathematical problems met in realistic situations, in particular as soon as we leave the linear framework, are not exactly solvable—either for an inherent impossibility or for our insufficient skills. Thus, Perturbation Theory is often the only way to approach realistic nonlinear systems.

It is implicit in the very nature of perturbation theory that it can only work once a problem which is both solvable—one also says “integrable”—and in some sense “near” to the original problem can be identified (it should be mentioned in this respect that the issue of “how near is near enough” is a delicate one).

Quite often, the integrable problem to be used as a starting point is a linear one—maybe obtained as the first-order expansion around a trivial or however known solution—and nonlinear corrections can be computed term by term via a recursive procedure based on expansion in a small parameter (usually denoted as  $\varepsilon$  by tradition); the point is that at each stage of this procedure one should only solve *linear* equations, so that the procedure can—at least in

principle—be carried over up to any desired order. In practice, one is limited by time, computational power, and the increasing dimension of the linear systems to be solved.

But limitations are not only due to the limits of the humans—or the computers—performing the actual computations: in fact, some delicate points arise when one considers the convergence of the  $\varepsilon$  series involved in the computations and in the expression of the solutions obtained by Perturbation Theory.

These points—i. e. the power of Perturbation Theory, its basic features and tools, and its limitations in particular with regard to convergence issues—are discussed in the article [▶ Perturbation Theory](#) by Gallavotti. This article also stresses the role which problems originating in Physics had in the development of Perturbation Theory; and this not only in historical terms (the computation of planetary orbits), but also in more recent times through the work of Poincaré first and then via Quantum Theory.

The modern setting of Perturbation Theory was laid down by Poincaré, and goes through the use of what is today known as *Poincaré normal forms*; these are a cornerstone of the whole theory and hence, implicitly or explicitly, of all the articles presented in this section of the Encyclopedia. But, they are also discussed in detail, together with their application, in the article [▶ Normal Forms in Perturbation Theory](#) by Broer.

The latter deals with the general problem, i. e. with evolution differential equations (Dynamical Systems) with no special structure; or, in applications originating from Physics or Engineering one is often dealing with systems that (within a certain approximation) preserve Energy and can be written in *Hamiltonian* form. In this case, as emphasized by Birkhoff, one can more efficiently consider perturbations of the Hamiltonian rather than of the equations of motion (the advantage originating in the fact that the Hamiltonian is a scalar function, while the equations of motion are a system of  $2n$  equations in  $2n$  dimensions). The normal form approach for Hamiltonian systems, and more generally Hamiltonian perturbation theory, is discussed in the article [▶ Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#) by Broer and Hanßmann. This also discusses the problem of transition—as some control parameter, often the Energy, is varied—from the regular behavior of the unperturbed system to the chaotic (“turbulent” if we deal with fluid motion) behavior displayed by many relevant Hamiltonian as well as non-Hamiltonian systems.

As mentioned above, in all the matters connected with Perturbation Theory and its applications, convergence issues play an extremely important role. They are discussed in the article [▶ Perturbative Expansions, Convergence of](#)



by Walcher, both in the general case and for Hamiltonian systems.

The interplay between perturbations—and more generally changes in some relevant parameter characterizing the system within a more general family of system—and qualitative (not only quantitative) changes in its behavior is of course of general interest not only in the “extreme” case of transition from integrable to chaotic behavior, but also when the qualitative change in the behavior of the system is somehow more moderate. Such a change is also known as a *bifurcation*. Albeit there is no article specifically devoted to these, the reader will note the concept of bifurcation appears in many, if not most, of the articles.

The behavior of a “generically perturbed” system depends on what is meant by “generically”. In particular if we deal with an unperturbed system which has some degree of symmetry, this may be an “accidental” feature—maybe due to the specially simple nature of integrable systems such as the one chosen as an unperturbed one—but might also correspond to a requirement by the very problem we are modeling; this is often the case when we deal with problems of physical or engineering origin, just because the fundamental equations of Physics have some degree of symmetry. The presence of symmetry can be quite helpful—e.g. reducing the effective degrees of freedom of a given problem—and should be taken into account in the perturbative expansion. Moreover, the perturbative expansions can be made to have some degree of symmetry which can be used in the solution of the resulting equations. These matters are discussed at length in the article [► Non-linear Dynamics, Symmetry and Perturbation Theory](#) in by Gaeta.

A special—but widely applicable and very interesting—framework for the occurrence of bifurcation is provided by systems exhibiting *parametric resonance*. This is, for example, the case for an ample class of coupled oscillator systems, which would per se suffice to guarantee the phenomenon is of special interest in applications, beside its theoretical appeal. The analysis of parametric resonance from the point of view of Perturbation Theory is discussed in the article [► Perturbation Analysis of Parametric Resonance](#) by Verhulst.

As mentioned above, the transition from fully regular (integrable) to chaotic behavior is discussed in general terms in the article [► Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#) by H. Broer and H. Hanßmann. However, quite remarkably, in some cases a perturbation will only moderately destroy the integrable behavior. This should be meant in the following sense: integrable behavior is characterized by the fact that whatever the initial conditions of the system, we are able to predict its behavior after an arbitrary long time. It may happen that albeit

this is not true, we are still able to predict either (a) the arbitrarily long time behavior for a dense subset of all the possible initial conditions; or (b) the exponentially long time behavior for a subset of full measure of possible initial conditions (usually, those “sufficiently near” to the exactly integrable case).

In the first case, the meaning of the statement is that any possible initial condition is “near” to an initial condition leading to an integrable-type behavior over all times (which does not imply its behavior will be near to integrable over arbitrary times, but only for sufficiently small—albeit this “small” could be extremely long on human scale—times). This kind of situation is investigated by the *KAM theory* (named after the initials of Kolmogorov, Arnold, and Moser), discussed in the article [► Kolmogorov–Arnold–Moser \(KAM\) Theory](#) by Chierchia.

In the second case, the statements about stable behavior are valid only for a finite (albeit exponentially long, hence again often extremely long on human scale) time, but apply to an open set of initial conditions. This approach was taken by Nekhoroshev, and is presently known—together with the results obtained in this direction—as *Nekhoroshev theory*; this is the subject of the article [► Nekhoroshev Theory](#) by Niederman.

As mentioned above, the problem of planetary motions was historically at the origin of Perturbation Theory, since Ancient Greece; actually more recent results, including those due to the work of Poincaré—and those embodied in KAM and Nekhoroshev theories—also have roots in Celestial Mechanics (albeit then being used in completely different fields, e.g. in the study of electron motion in a crystal). The application of Perturbation Theory in Celestial Mechanics is a very active field of research, and the subject of the article [► Perturbation Theory in Celestial Mechanics](#) by Celletti.

In this context, one often considers reduced problems where not all the planets are taken into account; this is the origin of the “three-body problem,” the three bodies being, for example the Sun, Jupiter, and the Earth; or the Earth, the Moon, and an artificial satellite. Much effort has been recently devoted to the study of special solutions for the N-body problem (that is, N bodies mutually attracting via potential forces) after the discovery of remarkable special solutions—termed “choreographies”—in which the bodies move along one or few common trajectories. This theory has not yet found applications in concrete physical systems or technology, but on the one hand these special solutions provide an organizing center for general nonlinear dynamics, and on the other the applicative potential of such collective motions (say in micro-devices) is rather

obvious. The N-body problem and these special solutions are discussed in the article ► [n-Body Problem and Choreographies](#) by Terracini.

The reader has probably noted that all these problems correspond to smooth *conservative* systems, or at least to perturbation of such systems. When this is relaxed—which is not appropriate in studying the motion of planetary objects, but may be appropriate in a number of contexts—the situation is both different and less well understood. The article ► [Perturbation of Systems with Nilpotent Real Part](#) by Gramchev studies perturbation of linear systems with a nilpotent linear part, which is the case for dissipative unperturbed systems. The article ► [Perturbation Theory for Non-smooth Systems](#) by Teixeira discusses the case of nonsmooth systems, which is the case—quite relevant in real-world engineering applications—of systems with impacts.

It was also mentioned that Quantum Theory was another major source of motivation for the development of Perturbation Theory, both historically and still in recent times. As for the latter, one should remark how a standard tool in quantum perturbation theory, i. e. the technique of *Feynman diagrams*—or more generally, diagrammatic expansions—was incorporated into classical perturbation theory only in relatively recent times. The use of diagrammatic expansions in classical perturbation theory is discussed in the article ► [Diagrammatic Methods in Classical Perturbation Theory](#) by Gentile.

Apart from this, knowledge of the perturbation-theoretic techniques developed in the framework of quantum theory—both in general and for the study of atoms and molecules—is of general interest, both as a source of inspiration for tackling problems in different contexts and for the intrinsic interest of microscopic systems; while well known to physicists, this theory is maybe less known to mathematicians and engineers. The articles provided in this section of the Encyclopedia can be an excellent entry point for those not familiar with this theory.

In the article ► [Perturbation Theory in Quantum Mechanics](#) by Picasso, Bracci, and D’Emilio, the general setting and results are described, together with some selected special topics; the role of symmetries—and hence degeneracies—within quantum perturbation theory is paramount and also discussed here.

The article ► [Perturbation Theory and Molecular Dynamics](#) by Panati focuses instead on the specific aspects of the perturbative approach to the quantum dynamics of molecules; this is a remarkable example of how taking into account the separation between slow and fast degrees of freedom allows one to deal with seemingly intractable problems.

A bridge between quantum and classical Perturbation Theory is provided by the *semiclassical* case, corresponding to taking into account the smallness of the energy scale set by Planck’s constant  $\hbar$  with respect to the energy scale involved in many (macro- or meso-scopic) problems. This is discussed in the article ► [Perturbation Theory, Semiclassical](#) by Sacchetti.

The quantum framework is also very interesting in connection with Bifurcation Theory; in this framework the “qualitative changes in the dynamics” which characterize bifurcations corresponds to qualitative changes in the spectrum. This in turn is related to *monodromy* on the mathematical side; and to the problem of an atom in crossed magnetic and electric fields on the physical side. These matters, strongly related to several of those mentioned above, are discussed in the article ► [Quantum Bifurcations](#) by Zhilinskii.

From the mathematical point of view, in the quantum case one deals with a Partial Differential Equation—the Schrödinger equation—rather than with a system of Ordinary Differential Equations (it should be noted that when dealing with the spectrum only, one is actually not requiring to study the full set of solutions to the concerned PDE).

Needless to say, this is not the only case where one has to deal with PDEs in the applications, continuum mechanics providing a classical framework where one is obliged to deal with PDEs. Rigorous results in Perturbation Theory for PDEs are not at the same level as for ODEs (and the insight provided by the quantum case is henceforth specially valuable); research in this direction is very active, and faces rather difficult problems despite the progresses obtained in recent years.

Some of these results, together with an overview of the field, are described in the article ► [Perturbation Theory for PDEs](#) by Bambusi. Quite appropriately, this article ends up stating that in several applied fields a sound understanding of PDEs rigorous Perturbation Theory would be relevant for applications, mentioning in particular the water wave problem, quantum mechanics, electromagnetic theory and magnetohydrodynamics, and elastodynamics.

This could also be a convenient way to conclude this Introduction, but in this case the reader would unavoidably remain with the impression that Perturbation Theory is mainly dealing with nonlinear problems originating in Physics or Engineering. While this is historically the origin of the most striking developments of the theory—and the realm in which it proved most successful—such characterization is by no means a built-in restriction.

Perturbation Theory can also deal effectively with problems originating in different fields and having a rather

different mathematical formulation, such as those arising in certain fields of Biology (beside fields where the mathematical formulation is anyway in terms of Dynamical Systems). This is shown concretely in the article ► [Perturbation of Equilibria in the Mathematical Theory of Evolution](#) by Sanchez; in this case the problem is formulated in terms of *Evolutionary Game Theory*. It should be noted that this is interesting not only for the intrinsic interest of the Darwin theory of Evolution, but also because Game Theory is increasingly used in rather diverse contexts.

Finally, I would like to warmly thank all the Authors for providing the remarkable articles making up this section of the *Encyclopedia*, as well as the Referees who checked them anonymously and in several cases gave suggestions leading to improvements.

## Perturbation Theory and Molecular Dynamics

GIANLUCA PANATI

Dipartimento di Matematica, Università di Roma  
“La Sapienza”, Roma, Italy

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Framework](#)

[The Leading Order Born–Oppenheimer Approximation](#)

[Beyond the Leading Order](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Adiabatic decoupling** In a complex system (either classical or quantum), the dynamical decoupling between the slow and the fast degrees of freedom.

**Adiabatic perturbation theory** A mathematical algorithm which exploits the adiabatic decoupling of degrees of freedom in order to provide an approximated (but yet accurate) description of the slow part of the dynamics. In the framework of QMD, it is used to approximately describe the dynamics of nuclei, the perturbative parameter  $\varepsilon$  being related to the small electron/nucleus mass ratio.

**Electronic structure problem** The problem consisting in computing, at fixed positions of the nuclei, the ener-

gies (eigenvalues) and eigenstates corresponding to the electrons. An approximate solution is usually obtained numerically.

**Molecular dynamics** The dynamics of the nuclei in a molecule. While a first insight in the problem can be obtained by using classical mechanics (Classical Molecular Dynamics), a complete picture requires quantum mechanics (Quantum Molecular Dynamics) ► [Perturbation Theory in Quantum Mechanics](#). This contribution focuses on the latter viewpoint.

### Definition of the Subject

In the framework of Quantum Mechanics the dynamics of a molecule is governed by the (time-dependent) Schrödinger equation, involving nuclei and electrons coupled through electromagnetic interactions. While the equation is mathematically well-posed, yielding the existence of a unique solution, the complexity of the problem makes the exact solution unattainable. Even for small molecules, the large number of degrees of freedom prevents from *direct* numerical simulation, making an approximation scheme necessary.

Indeed, one may exploit the smallness of the electron/nucleus mass ratio to introduce a convenient computational scheme leading to approximate solutions of the original time-dependent problem. In this article we review the standard approximation scheme (**dynamical Born–Oppenheimer approximation**) together with its ramifications and some recent generalizations, focusing on mathematically rigorous results.

The success of this approximation scheme is rooted in a clear separation of time-scales between the motion of electrons and nuclei. Such separation provides the prototypical example of **adiabatic decoupling** between the fast and the slow part of a quantum dynamics. More generally, adiabatic separation of time-scales plays a fundamental role in the understanding of complex system, with applications to a wide range of physical problems.

### Introduction

Through the discovery of the Schrödinger equation the theoretical physics and chemistry community attained a powerful tool for computing atomic spectra, either exactly or in perturbation expansion. Born and Oppenheimer [2] immediately strived for a more ambitious goal, namely to understand the excitation spectrum of molecules on the basis of the new wave mechanics. They accomplished to exploit the small electron/nucleus mass ratio as an expansion parameter, which then leads to

the **stationary** Born–Oppenheimer approximation. Since then it has become a standard and widely used tool in quantum chemistry, now supported by rigorous mathematical results [4,5,18,19,27].

Beyond molecular structure and excitation spectra, dynamical processes have gained in interest. Examples are the scattering of molecules, chemical reactions, or the decay of an excited state of a molecule through tunneling processes. Such problems require a **dynamical** version of the Born–Oppenheimer approximation, which is the topic of this article. At the leading order, the electronic energy at fixed positions of the nuclei serves as an effective potential between the nuclei. We call this the zeroth order Born–Oppenheimer approximation. The resulting effective Schrödinger equation can be used for both statical and dynamical purposes. The input is an electronic structure calculation, which for the purpose of our article we regard as given by other means.

While there are many physical and chemical properties of molecules explained by the zeroth order Born–Oppenheimer approximation, there are cases where higher order corrections are required. Famous examples are the dynamical Jahn–Teller effect and the dynamics of singled out nuclear degrees of freedom near the conical intersection of two energy surfaces. The first order Born–Oppenheimer approximation involves geometric phases, which are of great interest also in other domains of Quantum Mechanics ([41], ► [Quantum Bifurcations](#)).

Finally, we mention that some dynamical processes can be modeled as scattering problems. In such cases it is convenient to combine the Born–Oppenheimer scheme together with scattering theory, a topic which goes beyond the purpose of this contribution (see [28,29] and references therein).

A complete overview of the vast literature on the subject of the dynamical Born–Oppenheimer approximation is provided in [24].

## The Framework

We consider a molecule consisting of  $K$  nuclei, whose positions are denoted as  $x = (x_1, \dots, x_K) \in \mathbb{R}^{3K} =: X$ , and  $N$  electrons, with positions  $y = (y_1, \dots, y_N) \in \mathbb{R}^{3N} =: Y$ . The wavefunction of the molecule is therefore a square-integrable function  $\Psi$  depending on all these coordinates.

Molecular dynamics is described through the Schrödinger equation

$$i\hbar \frac{d}{ds} \Psi_s = H_{\text{mol}} \Psi_s, \quad (1)$$

where  $s$  denotes time measured in microscopic units and the Hamiltonian operator is given by

$$H_{\text{mol}} = - \sum_{k=1}^K \frac{\hbar^2}{2M_k} \Delta_{x_k} - \sum_{i=1}^N \frac{\hbar^2}{2m_e} \Delta_{y_i} + V_e(y) + V_n(x) + V_{\text{en}}(x, y). \quad (2)$$

Here  $\hbar$  is the Planck constant,  $m_e$  is the mass of the electron and  $M_k$  the mass of the  $k$ th nucleus, and the interaction terms are explicitly given by

$$V_n(x) = \sum_{k=1}^K \sum_{l \neq k}^K \frac{e^2 Z_k Z_l}{|x_k - x_l|}, \quad V_e(y) = \sum_{i=1}^N \sum_{j \neq i}^N \frac{e^2}{|y_i - y_j|},$$

and

$$V_{\text{en}}(x, y) = \sum_{k=1}^K \sum_{i=1}^N - \frac{e^2 Z_k}{|x_k - y_i|},$$

where  $eZ_k$ , for  $Z_k \in \mathbb{Z}$ , is the electric charge of the  $k$ th nucleus. In some cases, to obtain rigorous mathematical results one needs to slightly smear out the charge distribution of the nuclei. This is in agreement with the physical picture that nuclei are not point like but extended objects. Hereafter we will assume, for sake of a simpler notation, that all the nuclei have the same mass  $M$ . The subsequent discussion is still valid in the general case, with the appropriate choice of the adiabatic parameter indicated below.

As mentioned above, the large number of degrees of freedom makes convenient to elaborate an approximation scheme, exploiting the smallness of the parameter

$$\varepsilon := \sqrt{\frac{m_e}{M}} = 10^{-2} \dots 10^{-3}. \quad (3)$$

In the general case, one has to choose  $\varepsilon = \max\{\sqrt{m_e/M_k} : 1 \leq k \leq K\}$ .

By introducing atomic units ( $\hbar = 1, m_e = 1$ ) and making explicit the role of the **adiabatic parameter**  $\varepsilon$ , the Hamiltonian  $H_{\text{mol}}$  reads (up to a change of energy scale)

$$H_\varepsilon = - \sum_{k=1}^K \frac{\varepsilon^2}{2} \Delta_{x_k} + V_n(x) + \underbrace{\sum_{i=1}^N - \frac{1}{2} \Delta_{y_i} + V_e(y) + V_{\text{en}}(x, y)}_{H_{\text{el}}(x)}. \quad (4)$$

Notice that, for each fixed nuclei configuration  $x = (x_1, \dots, x_K) \in X$ , the operator  $H_{\text{el}}(x)$  is an operator acting

on the Hilbert space  $\mathcal{H}_{\text{el}}$  corresponding to the electrons alone.

If the kinetic energies of the nuclei and the electrons are comparable, as it happens in the vast majority of physical situations in view of energy equipartition, then the velocities scale as

$$|v_n| \approx \sqrt{\frac{m_e}{M}} |v_e| = \varepsilon |v_e|, \quad (5)$$

where  $v_n$  and  $v_e$  denotes respectively the typical velocity of nuclei and electrons. Therefore, in order to observe a non-trivial dynamics for the nuclei, one has to wait a microscopically long time, namely a time of order  $\mathcal{O}(\varepsilon^{-1})$ . This scaling fixes the **macroscopic time scale**  $t$ , together with the relation  $t = \varepsilon s$ , where  $s$  is the microscopic time appearing in Eq. (1). We are therefore interested in the behavior of the solutions of the equation

$$i\varepsilon \frac{d}{dt} \Psi(t) = H_\varepsilon \Psi(t) \quad (6)$$

in the limit  $\varepsilon \rightarrow 0$ .

We assume as an input a solution of the **electronic structure problem**, i. e. that for every fixed configuration of the nuclei  $x = (x_1, \dots, x_K)$  one knows the solution of the eigenvalue problem

$$H_e(x) \chi_j(x) = E_j(x) \chi_j(x), \quad (7)$$

with  $E_j(x) \in \mathbb{R}$  and  $\chi_j(x) \in \mathcal{H}_{\text{el}}$ . Since electrons are fermions, one has  $\mathcal{H}_{\text{el}} = S_a L^2(\mathbb{R}^{3N})$  with  $S_a$  projecting onto the antisymmetric wave functions. The eigenvectors in Eq. (7) are normalized as

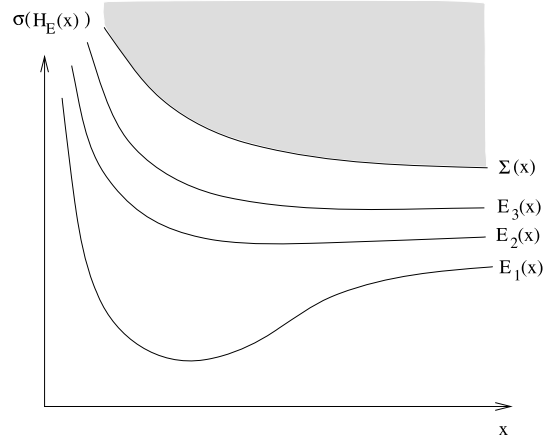
$$\langle \chi_j(x), \chi_\ell(x) \rangle_{\mathcal{H}_{\text{el}}} \equiv \int_Y \chi_j^*(x, y) \chi_\ell(x, y) dy = \delta_{j\ell}$$

with respect to the scalar product in  $\mathcal{H}_{\text{el}}$ . Note that the eigenvectors are determined only up to a phase  $\vartheta_j(x)$ . Generically, in addition to the bound states,  $H_e(x)$  has continuous spectrum. We label the eigenvalues in Eq. (7) as

$$E_1(x) \leq E_2(x) \leq \dots, \quad (8)$$

including multiplicity. The graph of  $E_j$  is called the  **$j$ th energy surface** or **energy band**, see Fig. 1.

Generically, in realistic examples such energy bands cross each other, and the possible structures of band crossing have been classified [22,35]. Figure 2 illustrates a realistic example of energy bands, showing in particular the typical conical intersection of two energy surfaces.



**Perturbation Theory and Molecular Dynamics, Figure 1**

A schematic representation of energy bands. At each fixed nuclei configuration  $x = (x_1, \dots, x_K)$  the electronic Hamiltonian  $H_{\text{el}}(x)$  exhibits point spectrum (corresponding to states with all the electrons bound to the nuclei) and continuous spectrum (corresponding to states in which one or more electrons are quasi-free, i. e. the molecule is ionized)

Let  $\psi(x)$  be a nucleonic wave function,  $\psi \in \mathcal{H}_n$ , the Hilbert space corresponding to the nuclei. For simplicity we take  $\mathcal{H}_n = L^2(X)$ , remembering that to impose the physically correct statistics for the nuclei requires extra considerations [33]. States  $\Psi$  of the molecules with the property that the electrons are precisely in the  $j$ th eigenstate are then of the form

$$\Psi(x, y) = \psi(x) \chi_j(x, y). \quad (9)$$

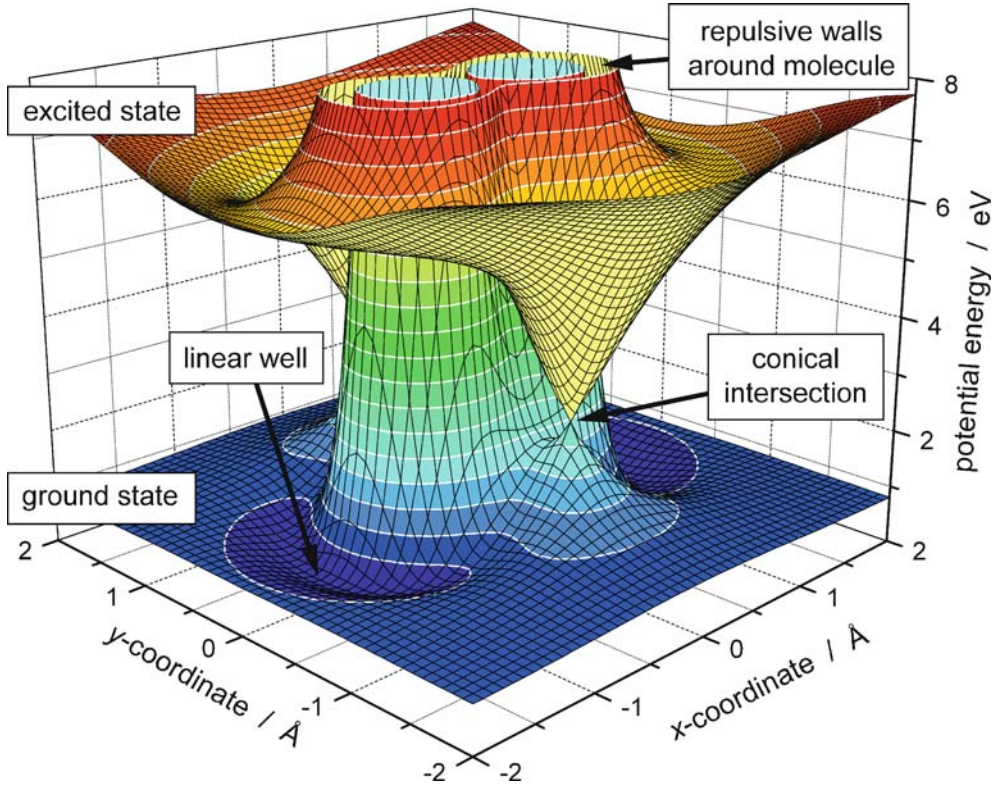
We can think of  $\Psi$  either as a wave function in the total Hilbert space  $\mathcal{H} = \mathcal{H}_n \otimes \mathcal{H}_{\text{el}}$ , or as a wave function for the electrons (i. e. an element of  $\mathcal{H}_{\text{el}}$ ) depending parametrically on  $x$ . In the common jargon, a state in the form Eq. (9) is said a state *concentrated on the  $j$ th band*. We denote as  $P_j$  the projector on the subspace consisting of states of the form Eq. (9); since the  $\{\chi_j(x)\}_{j \in \mathbb{N}}$  are orthonormal,  $P_j$  is indeed an orthogonal projection in  $\mathcal{H}_n \otimes \mathcal{H}_{\text{el}}$ .

### The Leading Order Born–Oppenheimer Approximation

We focus now on an a specific energy band, say  $E_n$ , assuming that it is globally isolated from the rest of the spectrum (the behavior of the wavefunction at the crossing points will be addressed later).

Under such assumption, a state  $\Psi_0$  which is initially concentrated on the  $n$ th band will stay localized in the same band up to errors of order  $\mathcal{O}(\varepsilon)$ : more specifically,





Perturbation Theory and Molecular Dynamics, Figure 2

The first two energy bands for the hydrogen quasi-molecule  $H_3$ , i.e. the system consisting of three protons and three electrons. The picture shows the restriction of such bands over a 2-dimensional subspace of the configuration space. Two hydrogen nuclei are located on the  $x$ -axis with a fixed separation of 1.044 Angstrom  $\text{\AA}$ , and the energy bands are plotted as a function of the relative position  $(x, y)$  of the third nucleus. Notice the conical intersection between the ground and the first excited state, which appears at equilateral triangular geometries. (© Courtesy of Eckart Wrede, Durham University. The plot is generated using the analytic representation of the  $H_3$  energy bands obtained in [40])

one shows that

$$\|(1 - P_n) e^{-iH_\varepsilon t/\varepsilon} P_n \Psi_0\|_{\mathcal{H}} = \mathcal{O}(\varepsilon). \quad (10)$$

The space  $\text{Ran } P_n$ , consisting of wavefunctions in the form Eq. (9), is usually called the **adiabatic subspace** corresponding to the  $n$ th band.

Since  $\text{Ran } P_n$  is approximately invariant under the dynamics, one may wonder whether there is a simple and convenient way to approximately describe the dynamics inside such subspace. Indeed, one may argue that for an initial state in  $\text{Ran } P_n$  the dynamics of the nuclei is governed by the reduced Hamiltonian

$$P_n H_\varepsilon P_n = -\frac{\varepsilon^2}{2} \sum_{k=1}^K \Delta_{x_k} + V_n(x) + E_n(x) + \mathcal{O}(\varepsilon) \quad (11)$$

acting in  $\text{Ran } P_n \cong \mathcal{H}_n = L^2(X)$ . The **dynamical Born-Oppenheimer approximation** consists, at the leading order,

in replacing the original Hamiltonian Eq. (4) by the Hamiltonian

$$H_{\text{BO}} = -\frac{\varepsilon^2}{2} \sum_{k=1}^K \Delta_{x_k} + V_n(x) + E_n(x) \quad (12)$$

acting in  $L^2(X)$ , getting thus an impressive dimensional reduction. In other words: let  $\Psi(t, x, y)$  be the solution of Eq. (6) with initial datum  $\Psi_0(x, y) = \varphi_0(x) \chi_n(x, y)$ ; then  $\Psi(t, x, y) = \varphi(t, x) \chi_n(x, y) + \mathcal{O}(\varepsilon)$  where  $\varphi(t, x)$  is the solution of the effective equation

$$i \frac{d}{dt} \varphi(t) = H_{\text{BO}} \varphi(t) \quad (13)$$

with initial datum  $\varphi_0$ .

To prove mathematically the previous claim, one has to bound the difference

$$(e^{-iH_\varepsilon t/\varepsilon} - e^{-iP_n H_\varepsilon P_n t/\varepsilon}) P_n.$$

A proof of this fact is not immediate as one might expect. Indeed, the Duhamel method (consisting essentially in rewriting a function as the integral of its derivative) yields

$$\begin{aligned} & (e^{-iH_\varepsilon t/\varepsilon} - e^{-iP_n H_\varepsilon P_n t/\varepsilon}) P_n \\ &= ie^{-iH_\varepsilon t/\varepsilon} \int_0^{t/\varepsilon} ds e^{iH_\varepsilon s} (P_n H_\varepsilon P_n - H_\varepsilon) e^{-iP_n H_\varepsilon P_n s} P_n \\ &= ie^{-iH_\varepsilon t/\varepsilon} \int_0^{t/\varepsilon} ds e^{iH_\varepsilon s} (P_n H_\varepsilon P_n - H_\varepsilon) P_n e^{-iP_n H_\varepsilon P_n s} \\ &= ie^{-iH_\varepsilon t/\varepsilon} \int_0^{t/\varepsilon} ds e^{iH_\varepsilon s} \underbrace{[P_n, H_\varepsilon] P_n}_{\mathcal{O}(\varepsilon)} e^{-iP_n H_\varepsilon P_n s}. \end{aligned}$$

The commutator appearing in the last line is estimated as

$$[P_n, H_\varepsilon] P_n = \left[ |\chi_n(x)\rangle \langle \chi_n(x)|, -\frac{\varepsilon^2}{2} \Delta_x \right] P_n = \mathcal{O}(\varepsilon), \quad (14)$$

but the integration interval diverges as  $\mathcal{O}(\varepsilon^{-1})$ . Thus the *naïve approach* fails. A rigorous proof has been provided in [39], elaborating on [26], exploiting the fact that the integral in Eq. (14) is, roughly speaking, an oscillatory operator integral. A more direct approach, based on the evolution of generalized Gaussian wavepackets, has been followed in the pioneering papers by Hagedorn [17,20].

### Beyond the Leading Order

The dynamics of a state initially concentrated on an isolated energy band is described, up to errors of order  $\mathcal{O}(\varepsilon)$ , by the Born–Oppenheimer dynamics Eq. (13). It is physically interesting to find an effective dynamics which approximates the original dynamics with an higher degree of accuracy. At first sight one might think that this goal can be simply reached by expanding the operator  $P_n H_\varepsilon P_n$  to the next order in  $\varepsilon$ . (Notice that the first term appearing in Eq. (11) does contribute as a term of  $\mathcal{O}(1)$ , since we are considering states such that the kinetic energy of the nuclei is not vanishing, i.e.  $\| -\varepsilon^2 \Delta_x \Psi \| = \mathcal{O}(1)$ , in agreement with the mentioned energy equipartition). However such *naïf* expansion has no physical meaning since it makes no sense to compute the operator appearing in Eq. (11) with greater accuracy if the space  $\text{Ran } P_n$  itself is invariant only up to terms of order  $\mathcal{O}(\varepsilon)$ .

To get a deeper insight in the problem, one has to investigate the origin of the  $\mathcal{O}(\varepsilon)$  term appearing in the Eq. (10): either (a) there is a part of the wavefunction of order  $\mathcal{O}(\varepsilon)$  which is scattered in all the directions in the Hilbert space, or (b) still there is a subspace invariant up

to smaller errors, which is however tilted with respect to  $\text{Ran } P_n$  by a term of order  $\mathcal{O}(\varepsilon)$ .

Therefore, two natural questions arise:

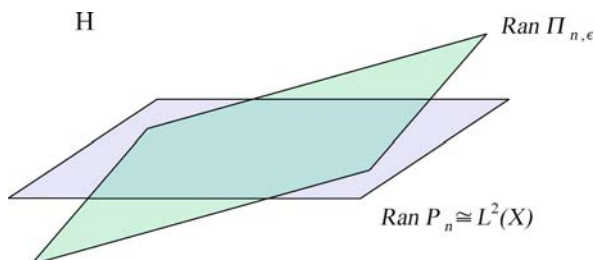
- Almost-invariant subspace:** is there a subspace of  $\mathcal{H} = \mathcal{H}_n \otimes \mathcal{H}_e$  which is invariant under the dynamics up to errors  $\varepsilon^N$ , for  $N > 1$ ?
- Intra-band dynamics:** in the affirmative case, is there any simple and convenient way to accurately describe the dynamics inside this subspace?

As for the first question, one may show that to any globally **isolated** energy band  $E_n$  corresponds a subspace of the Hilbert space which is almost-invariant under the dynamics. More precisely, one constructs an orthogonal projector  $\Pi_{n,\varepsilon} \in \mathcal{B}(\mathcal{H})$ , with  $\Pi_{n,\varepsilon} = P_n + \mathcal{O}(\varepsilon)$ , such that for any  $N \in \mathbb{N}$  there exists  $C_N$  such that

$$\begin{aligned} & \| (1 - \Pi_{n,\varepsilon}) e^{-iH_\varepsilon t/\varepsilon} \Pi_{n,\varepsilon} \Psi_0 \| \\ & \leq C_N \varepsilon^N (1 + |t|)(1 + \mathcal{E}) \|\Psi_0\|. \end{aligned} \quad (15)$$

Here  $\mathcal{E}$  denotes a cut-off on the kinetic energy of the nuclei, which corresponds to the physical assumption that the kinetic energies of nuclei and electrons are comparable. Equation (15) shows that if the molecule is initially in a state  $\Psi_0 \in \text{Ran } \Pi_{n,\varepsilon}$ , then after a macroscopic time  $t$  the molecule is in a state which is still in  $\text{Ran } \Pi_{n,\varepsilon}$  up to an error smaller than any power of  $\varepsilon$ , with the error scaling linearly with respect to time and to the kinetic energy cut-off. For this reason the space  $\text{Ran } \Pi_{n,\varepsilon}$  is called **superadiabatic subspace** or **almost-invariant subspace**.

We emphasize that the adiabatic decoupling, as formulated in Eq. (15), holds on a long time-scale, as opposed to the semiclassical limit which is known to hold on a time scale of order  $\mathcal{O}(\ln \varepsilon)$ . Indeed the adiabatic decoupling is a pure quantum phenomenon, conceptually and mathematically independent from the semiclassical limit.



**Perturbation Theory and Molecular Dynamics, Figure 3**

A schematic illustration of the superadiabatic subspace  $\text{Ran } \Pi_{n,\varepsilon}$ , tilted by a correction of order  $\varepsilon$  with respect to the usual adiabatic subspace  $\text{Ran } P_n$

The previous result is based on a long history of mathematical research, starting with pioneering ideas of Sjöstrand [3,9,31,32,34,37,38]. It has been formulated in the form above in [36].

As for question (ii), one has to face the problem that there is no natural identification between the super-adiabatic subspace  $\text{Ran } \Pi_{n,\varepsilon}$  and  $\mathcal{H}_n \cong L^2(X)$ , and therefore no evident reduction in the number of degrees of freedom. This difficulty can be circumvented by constructing a unitary operator which intertwines the previous two spaces, namely

$$U_{n,\varepsilon}: \text{Ran } \Pi_{n,\varepsilon} \rightarrow \mathcal{H}_n \cong L^2(X). \quad (16)$$

Notice that such a unitary operator is not unique. With the help of  $U_{n,\varepsilon}$  we can map the intraband dynamics to the nuclei Hilbert space, obtaining an effective Hamiltonian  $\hat{H}_{\text{eff},\varepsilon} := U_{n,\varepsilon} \Pi_{n,\varepsilon} H_\varepsilon \Pi_{n,\varepsilon} U_{n,\varepsilon}^{-1}$  acting in  $L^2(X)$ . It follows from Eq. (15) that for every  $N \in \mathbb{N}$  there exist  $C_N$  such that

$$\left\| \left( e^{-iH_\varepsilon t/\varepsilon} - U_{n,\varepsilon}^{-1} e^{-i\hat{H}_{\text{eff},\varepsilon} t/\varepsilon} U_{n,\varepsilon} \right) \Pi_{n,\varepsilon} \Psi_0 \right\|_{\mathcal{H}} \leq C_N \varepsilon^N (1 + |t|)(1 + \mathcal{E}) \|\Psi_0\|.$$

However, with an arbitrary choice of  $U_{n,\varepsilon}$  the effective Hamiltonian  $\hat{H}_{\text{eff},\varepsilon}$  does not appear simpler than  $\Pi_{n,\varepsilon} H_\varepsilon \Pi_{n,\varepsilon}$ . On the other side, the non-uniqueness of  $U_{n,\varepsilon}$  can be conveniently exploited to simplify the structure of the effective Hamiltonian. It has been proved in [36] that the unitary operator  $U_{n,\varepsilon}$  can be explicitly constructed in such a way that  $\hat{H}_{\text{eff},\varepsilon}$  has a simple structure, namely it is (close to) the  $\varepsilon$ -Weyl quantization of a function

$$H_{\text{eff},\varepsilon}: X \times \mathbb{R}^{3K} \rightarrow \mathbb{R}, \quad (q, p) \mapsto H_{\text{eff},\varepsilon}(q, p),$$

defined over the classical phase space. We recall that the  $\varepsilon$ -Weyl quantization maps a (smooth) function over  $X \times \mathbb{R}^{3K}$  into a (possibly unbounded) operator acting in  $L^2(X)$ . The correspondence is such that any function  $f(q)$  is mapped into the multiplication operator times  $f(x)$ , and any function  $g(p)$  is mapped into  $g(i\varepsilon \nabla_x)$ ; for a generic function  $f(q, p)$  the ordering ambiguity is fixed by choosing

$$e^{i\alpha \cdot q} e^{i\beta \cdot p} \mapsto e^{i(\alpha \cdot x + \beta \cdot (i\varepsilon \nabla_x))}.$$

For readers interested in the mathematical structure of Weyl quantization, we recommend [15].

Equipped with this terminology, we come back to the effective Hamiltonian. It turns out that, with the appropriate choice of  $U_{n,\varepsilon}$ ,  $\hat{H}_{\text{eff},\varepsilon}$  is the  $\varepsilon$ -Weyl quantization of the

function

$$H_{\text{eff},\varepsilon}(q, p) = h_0(q, p) + \varepsilon h_1(q, p) + \varepsilon^2 h_2(q, p) + \mathcal{O}(\varepsilon^3) \quad (17)$$

where

$$\begin{aligned} h_0(q, p) &= \frac{1}{2} p^2 + E_n(q) + V_n(q) \\ h_1(q, p) &= -ip \cdot \langle \chi_n(q), \nabla_q \chi_n(q) \rangle =: -p \cdot \mathcal{A}_n(q) \end{aligned} \quad (18)$$

and

$$\begin{aligned} h_2(q, p) &= \frac{1}{2} \mathcal{A}_n^2(q) + \frac{1}{2} \langle \nabla_q \chi_n(q), (1 - P_n(q)) \\ &\quad \cdot \nabla_q \chi_n(q) \rangle_{\mathcal{H}_{\text{el}}} \\ &\quad - \langle p \cdot \nabla_q \chi_n(q); \\ &\quad (H_{\text{el}}(q) - E_n(q))^{-1} (1 - P_n(q)) \\ &\quad \times p \cdot \nabla_q \chi_n(q) \rangle_{\mathcal{H}_{\text{el}}}. \end{aligned}$$

The Weyl quantization of  $h_0$  provides the leading order Born–Oppenheimer Hamiltonian Eq. (13). The term  $h_1$  has a geometric origin, involving the Berry connection  $\mathcal{A}_n(x)$ , a quantity appearing in a variety of adiabatic problems [41]; this term is responsible for the screening of magnetic fields in atoms [42]. Geometric effects in molecular systems (and more generally in adiabatic systems) are an active field of research [10,11,12], see ► [Quantum Bifurcations](#) and references therein. As for the second order correction  $h_2$ , the first term completes the square  $(p - \mathcal{A}_n(x))^2$  showing that the dynamics involves a covariant derivative; the second term is known as the Born–Huang term; the last term contains the reduced resolvent (i.e. the resolvent in the orthogonal complement of  $\text{Ran } P_n$ ) and is due to the fact that the superadiabatic subspace  $\text{Ran } \Pi_{n,\varepsilon}$  is tilted with respect to  $\text{Ran } P_n$ .

The third term in  $h_2$ , namely

$$\begin{aligned} \mathcal{M}(q, p) &= \langle p \cdot \nabla_q \chi_n(q), (H_{\text{el}}(q) - E_n(q))^{-1} \\ &\quad \times (1 - P_n(q)) p \cdot \nabla_q \chi_n(q) \rangle_{\mathcal{H}_{\text{el}}}, \end{aligned} \quad (19)$$

appeared firstly in [36], as a consequence of the rigorous adiabatic perturbation theory developed there. This term is responsible for an  $\mathcal{O}(\varepsilon^2)$ -correction to the effective mass of the nuclei. Indeed, since different quantization schemes differ by a term of order  $\mathcal{O}(\varepsilon)$ , we may replace the Weyl quantization with the simpler symmetric quantization, namely we consider

$$\begin{aligned} (\hat{\mathcal{M}}\psi)(x) &= \sum_{\ell,k=1}^{3K} \frac{1}{2} \left( m_{\ell k}(x) (-i\varepsilon \partial_{x_\ell}) (-i\varepsilon \partial_{x_k}) \right. \\ &\quad \left. + (-i\varepsilon \partial_{x_\ell}) (-i\varepsilon \partial_{x_k}) m_{\ell k}(x) \right) \psi(x), \end{aligned} \quad (20)$$

where  $m$  is the  $x$ -dependent matrix

$$m_{\ell k}(x) = \langle \partial_\ell \chi_n(x), (H_e(x) - E_n(x))^{-1} \times (1 - P_n(x)) \partial_k \chi_n(x) \rangle_{\mathcal{H}_{cl}}.$$

It is clear from Eq. (20) that this term induces a correction of order  $\mathcal{O}(\varepsilon^2)$  to the Laplacean, i. e. to the inertia of the nuclei.

Finally, we point out that the effective Hamiltonian  $\hat{H}_{\text{eff}, \varepsilon}$  can be conveniently truncated at any order in  $\varepsilon$ , getting corresponding errors in the effective quantum dynamics: if we pose

$$\hat{H}_{\text{eff}, \varepsilon} = \sum_{j=0}^N \varepsilon^j \hat{h}_j + \mathcal{O}(\varepsilon^{N+1}) =: \hat{h}_{(N), \varepsilon} + \mathcal{O}(\varepsilon^{N+1}), \quad (21)$$

then there exist a constant  $\tilde{C}_N$  such that

$$\left\| \left( e^{-iH_\varepsilon t/\varepsilon} - U_{n, \varepsilon}^{-1} e^{-i\hat{h}_{(N), \varepsilon} t/\varepsilon} U_{n, \varepsilon} \right) \Pi_{n, \varepsilon} \Psi_0 \right\|_{\mathcal{H}} \leq \tilde{C}_N \varepsilon^{N+1} (1 + |t|) (1 + \mathcal{E}) \|\Psi_0\|.$$

The determination of the effective Hamiltonian, here described following [36,40], has been investigated earlier in [31,41] with different but related techniques. The result in [36] is based on an iterative algorithm inspired by classical perturbation theory (see ► [Kolmogorov–Arnold–Moser \(KAM\) Theory](#) and ► [Normal Forms in Perturbation Theory](#)).

## Future Directions

Generically energy surfaces cross each other (see Fig. 2), and a globally isolated energy band is just a mathematical idealization. On the other side, if the initial datum  $\Psi_0(x, y) = \varphi_0(x) \chi_n(x, y)$  contains a nucleonic wavefunction  $\varphi_0$  localized far away from the crossing points, the adiabatic approximation is still valid, up to the time when the wavefunction becomes relevant in a neighborhood of radius  $\sqrt{\varepsilon}$  of the crossing points. This “hitting-time” can be estimated semiclassically, as done for example in [39].

When the wavefunction reaches the region around the crossing point a relevant part of it might undergo a transition to the other crossing band. (The simultaneous crossing of more than two bands is not generic, see [35], so we focus on the crossing of two bands). The understanding of the dynamics near a conical crossing is a very active field of research.

The first step is a convenient classification of the possible structures of band crossings. Since the early days of Quantum Mechanics [35], it has been realized that eigenvalue crossings occurs on submanifolds of various codimension, according to the symmetry of the problem. In

the case of a molecular Hamiltonian in the form Eq. (2), generic crossings of bands with the minimal multiplicity allowed by the symmetry group have been classified in [22].

The second step consist in an analysis of the propagation of the wavefunction near the conical crossing, assuming that the initial state is concentrated on a single band, say the  $n$ th band. A pioneering work [23] shows that the qualitative picture is the following: for crossings of codimension 1 the wavefunction follows, at the leading order, the analytic continuation of the  $n$ th band, as if there was no crossing. In the higher codimension case, a part of the wavefunction of order  $\mathcal{O}(1)$  undergoes a transition to the other band. More recently, propagation through conical crossings has been investigated with new techniques [6,7,8,13,14] opening the way to future research.

Alternatively, one may consider a family of energy bands which cross each other, but which are separated by an energy gap from the rest of the spectrum. Indeed, in a molecular collision or in excitations through a laser pulse only a few energy surfaces take part in the subsequent dynamics. Thus we take a set  $I$  of adjacent energy surfaces and call

$$P_I = \sum_{j \in I} P_j \quad (22)$$

the projection onto the relevant subspace (or subspace of physical interest). To ensure that other bands are not involved, we assume them to have a spectral gap of size  $a_{\text{gap}} > 0$  away from the energy surfaces in  $I$ , i. e.

$$\sup_{x \in X} |E_i(x) - E_j(x)| \geq a_{\text{gap}} \quad \text{for all } j \in I, i \in I^c. \quad (23)$$

Also the continuous spectrum is assumed to be at least  $a_{\text{gap}}$  away from the relevant energy surfaces. Under such assumption, the **multiband adiabatic theory** assures that the subspace  $\text{Ran } P_I$  is adiabatically protected against transitions, i. e.

$$\|(1 - P_I) e^{-iH_\varepsilon t/\varepsilon} P_I \Psi_0\|_{\mathcal{H}} = \mathcal{O}(\varepsilon).$$

Analogously to the case of a single band, one may construct the corresponding superadiabatic projector. The effective Hamiltonian  $\hat{H}_{\text{eff}, \varepsilon}$  corresponding to a family of  $m$  bands becomes, in this context, the  $\varepsilon$ -Weyl quantization of *matrix-valued* function over the classical phase space [36,40].

A deeper understanding of nuclear dynamics near conical crossings and a further developments of multiband adiabatic perturbation theory are, in the opinion of the author, two of the main directions for future research.



## Bibliography

### Primary Literature

- Berry MV, Lim R (1990) The Born–Oppenheimer electric gauge force is repulsive near degeneracies. *J Phys A* 23:L655–L657
- Born M, Oppenheimer R (1927) Zur Quantentheorie der Molekeln. *Ann Phys (Leipzig)* 84:457–484
- Brummelhuis R, Nourrigat J (1999) Scattering amplitude for Dirac operators. *Comm Partial Differ Equ* 24(1–2):377–394
- Combes JM (1977) The Born–Oppenheimer approximation. *Acta Phys Austriaca* 17:139–159
- Combes JM, Duclos P, Seiler R (1981) The Born–Oppenheimer approximation. In: Velo G, Wightman A (eds) *Rigorous Atomic and Molecular Physics*. Plenum, New York, pp 185–212
- de Verdière YC (2004) The level crossing problem in semi-classical analysis. II. The Hermitian case. *Ann Inst Fourier (Grenoble)* 54(5):1423–1441
- de Verdière YC, Lombardi M, Pollet C (1999) The microlocal Landau–Zener formula. *Ann Inst H Poincaré Phys Theor* 71:95–127
- de Verdière YC (2003) The level crossing problem in semi-classical analysis. I. The symmetric case. *Proceedings of the international conference in honor of Frédéric Pham (Nice, 2002)*. *Ann Inst Fourier (Grenoble)* 53(4):1023–1054
- Emmrich C, Weinstein A (1996) Geometry of the transport equation in multicomponent WKB approximations. *Commun Math Phys* 176:701–711
- Faure F, Zhilinskii BI (2000) Topological Chern indices in molecular spectra. *Phys Rev Lett* 85:960–963
- Faure F, Zhilinskii BI (2001) Topological properties of the Born–Oppenheimer approximation and implications for the exact spectrum. *Lett Math Phys* 55:219–238
- Faure F, Zhilinskii BI (2002) Topologically coupled energy bands in molecules. *Phys Lett* 302:242–252
- Fermanian–Kammerer C, Gérard P (2002) Mesures semi-classiques et croisement de modes. *Bull Soc Math France* 130:123–168
- Fermanian–Kammerer C, Lasser C (2003) Wigner measures and codimension 2 crossings. *J Math Phys* 44:507–527
- Folland GB (1989) *Harmonic analysis in phase space*. Princeton University Press, Princeton
- Hagedorn GA (1980) A time dependent Born–Oppenheimer approximation. *Commun Math Phys* 77:1–19
- Hagedorn GA (1986) High order corrections to the time-dependent Born–Oppenheimer approximation. I. Smooth potentials. *Ann Math (2)* 124(3):571–590
- Hagedorn GA (1987) High order corrections to the time-independent Born–Oppenheimer approximation I: Smooth potentials. *Ann Inst H Poincaré Sect. A* 47:1–16
- Hagedorn GA (1988) High order corrections to the time-independent Born–Oppenheimer approximation II: Diatomic Coulomb systems. *Comm Math Phys* 116:23–44
- Hagedorn GA (1988) High order corrections to the time-dependent Born–Oppenheimer approximation. II. Coulomb systems. *Comm Math Phys* 117(3):387–403
- Hagedorn GA (1989) Adiabatic expansions near eigenvalue crossings. *Ann Phys* 196:278–295
- Hagedorn GA (1992) Classification and normal forms for quantum mechanical eigenvalue crossings. *Méthodes semi-classiques*, vol 2 (Nantes, 1991). *Astérisque* 210(7):115–134
- Hagedorn GA (1994) Molecular propagation through electron energy level crossings. *Mem Amer Math Soc* 111:1–130
- Hagedorn GA, Joye A (2007) Mathematical analysis of Born–Oppenheimer approximations. *Spectral theory and mathematical physics: a Festschrift in honor of Barry Simon's 60th birthday*, In: *Proc Sympos Pure Math* 76, Part 1, Amer Math Soc, Providence, RI, pp 203–226
- Herrin J, Howland JS (1997) The Born–Oppenheimer approximation: straight-up and with a twist. *Rev Math Phys* 9:467–488
- Kato T (1950) On the adiabatic theorem of quantum mechanics. *Phys Soc Jap* 5:435–439
- Klein M, Martinez A, Seiler R, Wang XP (1992) On the Born–Oppenheimer expansion for polyatomic molecules. *Commun Math Phys* 143:607–639
- Klein M, Martinez A, Wang XP (1993) On the Born–Oppenheimer approximation of wave operators in molecular scattering theory. *Comm Math Phys* 152:73–95
- Klein M, Martinez A, Wang XP (1997) On the Born–Oppenheimer approximation of diatomic wave operators II. Singular potentials. *J Math Phys* 38:1373–1396
- Lasser C, Teufel S (2005) Propagation through conical crossings: an asymptotic transport equation and numerical experiments. *Commun Pure Appl Math* 58:1188–1230
- Littlejohn RG, Flynn WG (1991) Geometric phases in the asymptotic theory of coupled wave equations. *Phys Rev* 44:5239–5255
- Martinez A, Sordani V (2002) A general reduction scheme for the time-dependent Born–Oppenheimer approximation. *Comptes Rendus Acad Sci Paris* 334:185–188
- Mead CA, Truhlar DG (1979) On the determination of Born–Oppenheimer nuclear motion wave functions including complications due to conical intersections and identical nuclei. *J Chem Phys* 70:2284–2296
- Nenciu G, Sordani V (2004) Semiclassical limit for multistate Klein–Gordon systems: almost invariant subspaces and scattering theory. *J Math Phys* 45:3676–3696
- von Neumann J, Wigner EP (1929) Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. *Phys Z* 30:467–470
- Panati G, Spohn H, Teufel S (2003) Space-adiabatic perturbation theory. *Adv Theor Math Phys* 7:145–204
- Sjöstrand J (1993) Projecteurs adiabatiques du point de vue pseudodifférentiel. *Comptes Rendus Acad Sci Paris, Série I* 317:217–220
- Sordani V (2003) Reduction scheme for semiclassical operator-valued Schrödinger type equation and application to scattering. *Comm Partial Differ Equ* 28(7–8):1221–1236
- Spohn H, Teufel S (2001) Adiabatic decoupling and time-dependent Born–Oppenheimer theory. *Commun Math Phys* 224:113–132
- Varandas AJC, Brown FB, Mead CA, Truhlar DG, Blais NC (1987) A double many-body expansion of the two lowest-energy potential surfaces and nonadiabatic coupling for H<sub>3</sub>. *J Chem Phys* 86:6258–6269
- Weigert S, Littlejohn RG (1993) Diagonalization of multicomponent wave equations with a Born–Oppenheimer example. *Phys Rev A* 47:3506–3512
- Yin L, Mead CA (1994) Magnetic screening of nuclei by electrons as an effect of geometric vector potential. *J Chem Phys* 100:8125–8131



## Books and Reviews

- Bohm A, Mostafazadeh A, Koizumi A, Niu Q, Zwanziger J (2003) The geometric phase in quantum systems. Texts and monographs in physics. Springer, Heidelberg
- Teufel S (2003) Adiabatic perturbation theory in quantum dynamics. Lecture notes in mathematics, vol 1821. Springer, Berlin

## Perturbation Theory for Non-smooth Systems

MARCO ANTÔNIO TEIXEIRA

Department of Mathematics, Universidade Estadual de Campinas, Campinas, Brazil

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Preliminaries  
Vector Fields near the Boundary  
Generic Bifurcation  
Singular Perturbation Problem in 2D  
Future Directions  
Bibliography

### Glossary

**Non-smooth dynamical system** Systems derived from ordinary differential equations when the non-uniqueness of solutions is allowed. In this article we deal with discontinuous vector fields in  $R^n$  where the discontinuities are concentrated in a codimension-one surface.

**Bifurcation** In a  $k$ -parameter family of systems, a bifurcation is a parameter value at which the phase portrait is not structurally stable.

**Typical singularity** Are points on the discontinuity set where the orbits of the system through them must be distinguished.

### Definition of the Subject

In this article we survey some qualitative and geometric aspects of non-smooth dynamical systems theory. Our goal is to provide an overview of the state of the art on the theory of contact between a vector field and a manifold, and on discontinuous vector fields and their perturbations. We also establish a bridge between two-dimensional non-smooth systems and the geometric singular perturbation theory. Non-smooth dynamical systems is a subject that has been developing at a very fast pace in recent years due

to various factors: its mathematical beauty, its strong relationship with other branches of science and the challenge in establishing reasonable and consistent definitions and conventions. It has become certainly one of the common frontiers between mathematics and physics/engineering. We mention that certain phenomena in control systems, impact in mechanical systems and nonlinear oscillations are the main sources of motivation for our study concerning the dynamics of those systems that emerge from differential equations with discontinuous right-hand sides. We understand that non-smooth systems are driven by applications and they play an intrinsic role in a wide range of technological areas.

### Introduction

The purpose of this article is to present some aspects of the geometric theory of a class of non-smooth systems. Our main concern is to bring the theory into the domain of geometry and topology in a comprehensive mathematical manner.

Since this is an impossible task, we do not attempt to touch upon all sides of this subject in one article. We focus on exploring the local behavior of systems around typical singularities. The first task is to describe a generic persistence of a local theory (structural stability and bifurcation) for discontinuous systems mainly in the two- and three-dimensional cases. Afterwards we present some striking features and results of the regularization process of two-dimensional discontinuous systems in the framework developed by Sotomayor and Teixeira in [44] and establish a bridge between those systems and the fundamental role played by the Geometric Singular Perturbation Theory (GSPT). This transition was introduced in [10] and we reproduce here its main features in the two-dimensional case. For an introductory reading on the methods of geometric singular perturbation theory we refer to [16,18,30]. In Sect. “Definition of the Subject” we introduce the setting of this article. In Sect. “Introduction” we survey the state of the art of the contact between a vector field and a manifold. The results contained in this section are crucial for the development of our approach. In Sect. “Preliminaries” we discuss the classification of typical singularities of non-smooth vector fields. The study of non-smooth systems, via GSPT, is presented in Sect. “Vector Fields near the Boundary”. In Sect. “Generic Bifurcation” some theoretical open problems are presented.

One aspect of the qualitative point of view is the problem of structural stability, the most comprehensive of many different notions of stability. This theme was studied in 1937 by Andronov–Pontryagin (see [3]). This prob-

lem is of obvious importance, since in practice one obtains a lot of qualitative information not only on a fixed system but also on its nearby systems.

We deal with non-smooth vector fields in  $R^{n+1}$  having a codimension-one submanifold  $M$  as its discontinuity set. The scheme in this work toward a systematic classification of typical singularities of non-smooth systems follows the ideas developed by Sotomayor–Teixeira in [43] where the problem of contact between a vector field and the boundary of a manifold was discussed. Our approach intends to be self-contained and is accompanied by an extensive bibliography. We will try to focus here on areas that are complementary to some recent reviews made elsewhere.

The concept of structural stability in the space of non-smooth vector fields is based on the following definition:

**Definition 1** Two vector fields  $Z$  and  $\tilde{Z}$  are  $C^0$  equivalent if there is an  $M$ -invariant homeomorphism  $h: R^{n+1} \rightarrow R^{n+1}$  that sends orbits of  $Z$  to orbits of  $\tilde{Z}$ .

A general discussion is presented to study certain unstable non-smooth vector fields within a generic context. The framework in which we shall pursue these unstable systems is sometimes called generic bifurcation theory. In [3] the concept of  $k$ th-order structural stability is also presented; in a local approach such setting gives rise to the notion of a codimension- $k$  singularity. In studies of classical dynamical systems, normal form theory has been well accepted as a powerful tool in studying the local theory (see [6]). Observe that, so far, bifurcation and normal form theories for non-smooth vector fields have not been extensively studied in a systematic way.

Control Theory is a natural source of mathematical models of these systems (see, for instance, [4,8,20,41,45]). Interesting problems concerning discontinuous systems can be formulated in systems with hysteresis ([41]), economics ([23,25]) and biology ([7]). It is worth mentioning that in [5] a class of relay systems in  $R^n$  is discussed. They have the form:

$$\dot{X} = Ax + \operatorname{sgn}(x_1)k$$

where  $x = (x_1, x_2, \dots, x_n)$ ,  $A \in M_R(n, n)$  and  $k = (k_1, k_2, \dots, k_n)$  is a constant vector in  $R^n$ . In [28,29] the generic singularities of reversible relay systems in 4D were classified. In [54] some properties of non-smooth dynamics are discussed in order to understand some phenomena that arise in chattering control. We mention the presence of chaotic behavior in some non-smooth systems (see for example [12]). It is worthwhile to cite [17], where the main problem in the classical calculus of variations was carried out to study discontinuous Hamiltonian vector fields.

We refer to [14] for a comprehensive text involving non-smooth systems which includes many models and applications. In particular motivating models of several non-smooth dynamical systems arising in the occurrence of impacting motion in mechanical systems, switchings in electronic systems and hybrid dynamics in control systems are presented together with an extensive literature on impact oscillators which we do not attempt to survey here. For further reading on some mathematical aspects of this subject we recommend [11] and references therein. A setting of general aspects of non-smooth systems can be found also in [35] and references therein. Our discussion does not focus on continuous but rather on non-smooth dynamical systems and we are aware that the interest in this subject goes beyond the approach adopted here.

The author wishes to thank R. Garcia, T.M. Seara and J. Sotomayor for many helpful conversations.

## Preliminaries

Now we introduce some of the terminology, basic concepts and some results that will be used in the sequel.

**Definition 2** Two vector fields  $Z$  and  $\tilde{Z}$  on  $R^n$  with  $Z(0) = \tilde{Z}(0)$  are *germ-equivalent* if they coincide on some neighborhood  $V$  of 0.

The equivalent classes for this equivalence are called germs of vector fields. In the same way as defined above, we may define germs of functions. For simplicity we are considering the germ notation and we will not distinguish a germ of a function and any one of its representatives. So, for example, the notation  $h: R^n, 0 \rightarrow R$  means that the  $h$  is a germ of a function defined in a neighborhood of 0 in  $R^n$ . Refer to [15] for a brief and nice introduction of the concepts of *germ* and *k-jet* of functions.

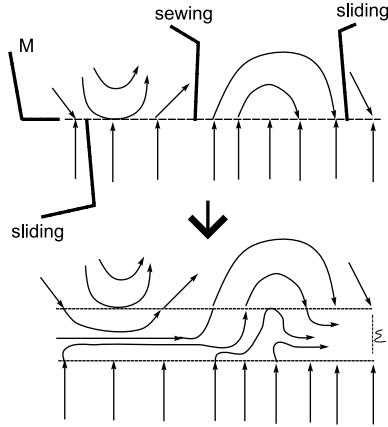
## Discontinuous Systems

Let  $M = h^{-1}(0)$ , where  $h$  is (a germ of) a smooth function  $h: R^{n+1}, 0 \rightarrow R$  having  $0 \in R$  as its regular value. We assume that  $0 \in M$ .

Designate by  $\chi(n+1)$  the space of all germs of  $C^r$  vector fields on  $R^{n+1}$  at 0 endowed with the  $C^r$ -topology with  $r > 1$  and large enough for our purposes. Call  $\Omega(n+1)$  the space of all germs of vector fields  $Z$  in  $R^{n+1}, 0$  such that

$$Z(q) = \begin{cases} X(q), & \text{for } h(q) > 0, \\ Y(q), & \text{for } h(q) < 0, \end{cases} \quad (1)$$

The above field is denoted by  $Z = (X, Y)$ . So we are considering  $\Omega(n+1) = \chi(n+1) \times \chi(n+1)$  endowed with the product topology.



**Perturbation Theory for Non-smooth Systems, Figure 1**  
A discontinuous system and its regularization

**Definition 3** We say that  $Z \in \Omega(n+1)$  is structurally stable if there exists a neighborhood  $U$  of  $Z$  in  $\Omega(n+1)$  such that every  $\tilde{Z} \in U$  is  $C^0$ -equivalent with  $Z$ .

To define the orbit solutions of  $Z$  on the switching surface  $M$  we take a pragmatic approach. In a well characterized open set  $O$  of  $M$  (described below) the solution of  $Z$  through a point  $p \in O$  obeys the Filippov rules and on  $M - O$  we accept it to be multivalued. Roughly speaking, as we are interested in studying the structural stability in  $\Omega(n+1)$  it is convenient to take into account all the leaves of the foliation in  $R^{n+1}$  generated by the orbits of  $Z$  (and also the orbits of  $X$  and  $Y$ ) passing through  $p \in M$ . (see Fig. 1)

The trajectories of  $Z$  are the solutions of the autonomous differential system  $\dot{q} = Z(q)$ .

In what follows we illustrate our terminology by presenting a simplified model that is found in the classical electromagnetism theory (see for instance [26]):

$$\ddot{x} - \ddot{x} + \alpha \operatorname{sign} x = 0.$$

with  $\alpha > 0$ .

So this system can be expressed by the following objects:  $h(x, y, z) = x$  and  $Z = (X, Y)$  with  $X(x, y, z) = (y, z, z + \alpha)$  and  $Y(x, y, z) = (y, z, z - \alpha)$ .

For each  $X \in \chi(n+1)$  we define the smooth function  $Xh: R^{n+1} \rightarrow R$  given by  $Xh = X \cdot \nabla h$  where  $\cdot$  is the canonical scalar product in  $R^{n+1}$ .

We distinguish the following regions on the discontinuity set  $M$ :

- (i)  $M_1$  is the *sewing region* that is represented by  $h = 0$  and  $(Xh)(Yh) > 0$ ;

- (ii)  $M_2$  is the *escaping region* that is represented by  $h = 0$ ,  $(Xh) > 0$  and  $(Yh) < 0$ ;
- (iii)  $M_3$  is the *sliding region* that is represented by  $h = 0$ ,  $(Xh) < 0$  and  $(Yh) > 0$ .

We set  $\mathcal{O} = \bigcup_{i=1,2,3} M_i$ .

Consider  $Z = (X, Y) \in \Omega(n+1)$  and  $p \in M_3$ . In this case, following Filippov's convention, the solution  $\gamma(t)$  of  $Z$  through  $p$  follows, for  $t \geq 0$ , the orbit of a vector field tangent to  $M$ . Such system is called *sliding vector field* associated with  $Z$  and it will be defined below.

**Definition 4** The sliding vector field associated to  $Z = (X, Y)$  is the smooth vector field  $Z^s$  tangent to  $M$  and defined at  $q \in M_3$  by  $Z^s(q) = m - q$  with  $m$  being the point where the segment joining  $q + X(q)$  and  $q + Y(q)$  is tangent to  $M$ .

It is clear that if  $q \in M_3$  then  $q \in M_2$  for  $-Z$  and then we define the *escaping vector field* on  $M_2$  associated with  $Z$  by  $Z^e = -(-Z)^s$ . In what follows we use the notation  $Z^M$  for both cases.

We recall that sometimes  $Z^M$  is defined in an open region  $U$  with boundary. In this case it can be  $C^r$  extended to a full neighborhood of  $p \in \partial U$  in  $M$ .

When the vectors  $X(p)$  and  $Y(p)$ , with  $p \in M_2 \cup M_3$  are linearly dependent then  $Z^M(p) = 0$ . In this case we say that  $p$  is a simple singularity of  $Z$ . The other singularities of  $Z$  are concentrated outside the set  $O$ .

We finish this subsection with a three-dimensional example:

Let  $Z = (X, Y) \in \Omega(3)$  with  $h(x, y, z) = z$ ,  $X = (1, 0, x)$  and  $Y = (0, 1, y)$ . The system determines four quadrants around 0, bounded by  $\tau_X = \{x = 0\}$  and  $\tau_Y = \{y = 0\}$ . They are:  $Q_1^+ = \{x > 0, y > 0\}$ ,  $Q_1^- = \{x < 0, y < 0\}$ ,  $Q_2 = \{x < 0, y > 0\}$  (sliding region) and  $Q_3 = \{x > 0, y < 0\}$  (escaping region). Observe that  $M_1 = Q_1^+ \cup Q_1^-$ .

The sliding vector field defined in  $Q_2$  is expressed by:

$$Z^s(x, y, z) = (y - x)^{-1} \left( x + y, \frac{y + x}{8}, 0 \right).$$

Such a system is (in  $Q_2$ ) equivalent to  $G(x, y, z) = (x + y, \frac{y+x}{8}, 0)$ . In our terminology we consider  $G$  a smooth extension of  $Z^s$ , that is defined in a whole neighborhood of 0. It is worthwhile to say that  $G$  is in fact a system which is equivalent to the original system in  $Q_2$ .

In [50] a generic classification of one-parameter families of sliding vector fields is presented.

### Singular Perturbation Problem

A singular perturbation problem is expressed by a differential equation  $z' = \alpha(z, \varepsilon)$  (refer to [16,18,30]) where  $z \in R^{n+m}$ ,  $\varepsilon$  is a small non-negative real number and  $\alpha$  is a  $C^\infty$  mapping.

Let  $z = (x, y) \in R^{n+m}$  and  $f: R^{m+n} \rightarrow R^m$ ,  $g: R^{m+n} \rightarrow R^n$  be smooth mappings. We deal with equations that may be written in the form

$$\begin{cases} x' = f(x, y, \varepsilon) \\ y' = \varepsilon g(x, y, \varepsilon) \end{cases} \quad x = x(\tau), y = y(\tau). \quad (2)$$

An interesting model of such systems can be obtained from the singular van der Pol's equation

$$\varepsilon x'' + (x^2 + x)x' + x - a = 0. \quad (3)$$

The main trick in the geometric singular perturbation (GSP) is to consider the family (2) in addition to the family

$$\begin{cases} \varepsilon \dot{x} = f(x, y, \varepsilon) \\ \dot{y} = g(x, y, \varepsilon) \end{cases} \quad x = x(t), y = y(t) \quad (4)$$

obtained after the time rescaling  $t = \varepsilon \tau$ .

Equation (2) is called the *fast system* and (4) the *slow system*. Observe that for  $\varepsilon > 0$  the phase portrait of fast and slow systems coincide.

For  $\varepsilon = 0$ , let  $S$  be the set of all singular points of (2). We call  $S$  the slow manifold of the singular perturbation problem and it is important to notice that Eq. (4) defines a dynamical system on  $S$  called the *reduced problem*.

Combining results on the dynamics of these two limiting problems (2) and (4), with  $\varepsilon = 0$ , one obtains information on the dynamics for small values of  $\varepsilon$ . In fact, such techniques can be exploited to formally construct approximated solutions on pieces of curves that satisfy some limiting version of the original equation as  $\varepsilon$  goes to zero.

**Definition 5** Let  $A, B \subset R^{n+m}$  be compact sets. The Hausdorff distance between  $A$  and  $B$  is  $D(A, B) = \max_{z_1 \in A, z_2 \in B} \{d(z_1, B), d(z_2, A)\}$ .

The main question in GSP-theory is to exhibit conditions under which a singular orbit can be approximated by regular orbits for  $\varepsilon \downarrow 0$ , with respect to the Hausdorff distance.

### Regularization Process

An approximation of the discontinuous vector field  $Z = (X, Y)$  by a one-parameter family of continuous vector fields will be called a regularization of  $Z$ . In [44], Sotomayor and Teixeira introduced the regularization procedure of a discontinuous vector field. A transition function

is used to average  $X$  and  $Y$  in order to get a family of continuous vector fields that approximates the discontinuous one. Figure 1 gives a clear illustration of the regularization process.

Let  $Z = (X, Y) \in \Omega(n+1)$ .

**Definition 6** A  $C^\infty$  function  $\varphi: R \rightarrow R$  is a transition function if  $\varphi(x) = -1$  for  $x \leq -1$ ,  $\varphi(x) = 1$  for  $x \geq 1$  and  $\varphi'(x) > 0$  if  $x \in (-1, 1)$ . The  $\phi$ -regularization of  $Z = (X, Y)$  is the one-parameter family  $X_\varepsilon \in C^r$  given by

$$Z_\varepsilon(q) = \left( \frac{1}{2} + \frac{\varphi_\varepsilon(h(q))}{2} \right) X(q) + \left( \frac{1}{2} - \frac{\varphi_\varepsilon(h(q))}{2} \right) Y(q). \quad (5)$$

with  $h$  given in the above Subject. “Discontinuous Systems” and  $\varphi_\varepsilon(x) = \varphi(x/\varepsilon)$ , for  $\varepsilon > 0$ .

As already said before, a point in the phase space which moves on an orbit of  $Z$  crosses  $M$  when it reaches the region  $M_1$ . Solutions of  $Z$  through points of  $M_3$ , will remain in  $M$  in forward time. Analogously, solutions of  $Z$  through points of  $M_2$  will remain in  $M$  in backward time. In [34,44] such conventions are justified by the regularization method in dimensions two and three respectively.

### Vector Fields near the Boundary

In this section we discuss the behavior of smooth vector fields in  $R^{n+1}$  relative to a codimension-one submanifold (say, the above defined  $M$ ). We base our approach on the concepts and results contained in [43,53]. The principal advantage of this setting is that the generic contact between a smooth vector field and  $M$  can often be easily recognized. As an application the typical singularities of a discontinuous system can be further classified in a straightforward way.

We say that  $X, Y \in \chi(n+1)$  are  $M$ -equivalent if there exists an  $M$ -preserving homeomorphism  $h: R^{n+1}, 0 \rightarrow R^{n+1}, 0$  that sends orbits of  $X$  into orbits of  $Y$ . In this way we get the concept of  $M$ -structural stability in  $\chi(n+1)$ .

We call  $\Gamma_0(n+1)$  the set of elements  $X$  in  $\chi(n+1)$  satisfying one of the following conditions:

- 0)  $Xh(0) \neq 0$  (0 is a regular point of  $X$ ). In this case  $X$  is transversal to  $M$  at 0.
- 1)  $Xh(0) = 0$  and  $X^2h(0) \neq 0$  (0 is a 2-fold point of  $X$ );
- 2)  $Xh(0) = X^2h(0) = 0$ ,  $X^3h(0) \neq 0$  and the set  $\{Dh(0), DXh(0), DX^2h(0)\}$  is linearly independent (0 is a cusp point of  $X$ );
- ...
- n)  $Xh(0) = X^2h(0) = \dots = X^n h(0) = 0$  and  $X^{n+1}h(0) \neq 0$ . Moreover the set  $\{Dh(0), DXh(0),$

$DX^2h(0), \dots, DX^n h(0)\}$  is linearly independent, and 0 is a regular point of the mapping  $Xh|_M$ .

We say that 0 is an  $M$ -singularity of  $X$  if  $h(0) = Xh(0) = 0$ . It is a *codimension-zero*  $M$ -singularity provided that  $X \in \Gamma_0(n+1)$ .

We know that  $\Gamma_0(n+1)$  is an open and dense set in  $\chi(n+1)$  and it coincides with the  $M$ -structurally stable vector fields in  $\chi(n+1)$  (see [53]).

Denote by  $\tau_X \subset M$  the  $M$ -singular set of  $X \in \chi(n+1)$ ; this set is represented by the equations  $h = Xh = 0$ . It is worthwhile to point out that, generically, all two-folds constitute an open and dense subset of  $\tau_X$ . Observe that if  $X(0) = 0$  then  $X \notin \Gamma_0(n+1)$ .

The  $M$ -bifurcation set is represented by  $\chi_1(n+1) = \chi(n+1) - \Gamma_0(n+1)$ .

Vishik in [53] exhibited the normal forms of a *codimension-zero*  $M$ -singularity. They are:

I) Straightened vector field

$$X = (1, 0, \dots, 0)$$

and

$$h(x) = x_1^{k+1} + x_2 x_1^{k-1} + x_3 x_1^{k-2} + \dots + x_{k+1}, \quad k = 0, 1, \dots, n$$

or

II) Straightened boundary

$$h(x) = x_1$$

and

$$X(x) = (x_2, x_3, \dots, x_k, 1, 0, 0, \dots, 0)$$

We now discuss an important interaction between vector fields near  $M$  and singularities of mapping theory. We discuss how singularity-theoretic techniques help the understanding of the dynamics of our systems.

We outline this setting, which will be very useful in the sequel. The starting point is the following construction.

### A Construction

Let  $X \in \chi(n+1)$ . Consider a coordinate system  $x = (x_1, x_2, \dots, x_{n+1})$  in  $R^{n+1}$ , 0 such that

$$M = \{x_1 = 0\}$$

and

$$X = (X^1, X^2, \dots, X^{n+1})$$

Assume that  $X(0) \neq 0$  and  $X^1(0) = 0$ . Let  $N_0$  be any transversal section to  $X$  at 0.

By the implicit function theorem, we derive that:

for each  $p \in M$ , 0 there exists a unique  $t = t(p)$  in  $R$ , 0 such that the orbit-solution  $t \mapsto \gamma(p, t)$  of  $X$  through  $p$  meets  $N_0$  at a point  $\tilde{p} = \gamma(p, t(p))$ .

We define the smooth mapping  $\rho_X: R^n, 0 \rightarrow R^n, 0$  by  $\rho_X(p) = \tilde{p}$ . This mapping is a powerful tool in the study of vector fields around the boundary of a manifold (refer to [21,42,43,46,53]). We observe that  $\tau_X$  coincides with the singular set of  $\rho_X$ .

The late construction implements the following method. If we are interested in finding an equivalence between two vector fields which preserve  $M$ , then the problem can be sometimes reduced to finding an equivalence between  $\rho_X$  and  $\rho_Y$  in the sense of singularities of mappings.

We recall that when 0 is a fold  $M$ -singularity of  $X$  then associated to the fold mapping  $\rho_X$  there is the symmetric diffeomorphism  $\beta_X$  that satisfies  $\rho_X \circ \beta_X = \rho_X$ .

Given  $Z = (X, Y) \in \Omega(n+1)$  such that  $\rho_X$  and  $\rho_Y$  are fold mappings with  $X^2h(0) < 0$  and  $Y^2h(0) > 0$  then the composition of the associated symmetric mappings  $\beta_X$  and  $\beta_Y$  provides a first return mapping  $\beta_Z$  associated to  $Z$  and  $M$ . This situation is usually called a *distinguished fold-fold* singularity, and the mapping  $\beta_Z$  plays a fundamental role in the study of the dynamics of  $Z$ .

### Codimension-one $M$ -Singularity in Dimensions Two and Three

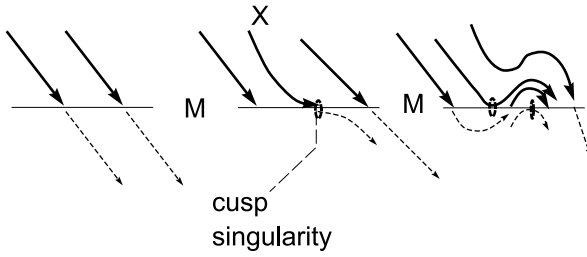
**Case  $n = 1$**  In this case the unique codimension-zero  $M$ -singularity is a fold point in  $R^2, 0$ . The codimension-one  $M$ -singularities are represented by the subset  $\Gamma_1(2)$  of  $\chi_1(2)$  and it is defined as follows.

**Definition 7** A codimension-one  $M$ -singularity of  $X \in \Gamma_1(2)$  is either a *cusp* singularity or an  $M$ -hyperbolic critical point  $p$  in  $M$  of the vector field  $X$ . A cusp singularity (illustrated in Fig. 2) is characterized by  $Xh(p) = X^2h(p) = 0$ ,  $X^3h(p) \neq 0$ . In the second case this means that  $p$  is a hyperbolic critical point (illustrated in Fig. 3) of  $X$  with distinct eigenvalues and with invariant manifolds (stable, unstable and strong stable and strong unstable) transversal to  $M$ .

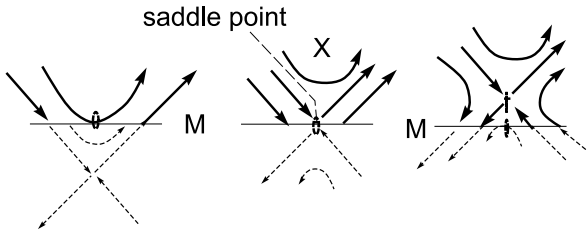
In this subsection we consider a coordinate system in  $R^2, 0$  such that  $h(x, y) = y$ .

The next result was proved in [46]. It presents the normal forms of the codimension-one singularities defined above.





**Perturbation Theory for Non-smooth Systems, Figure 2**  
The cusp singularity and its unfolding



**Perturbation Theory for Non-smooth Systems, Figure 3**  
The saddle point in the boundary and its unfolding

**Theorem 8** Let  $X \in \chi_1(2)$ . The vector field  $X$  is  $M$ -structurally stable relative to  $\chi_1(2)$  if and only if  $X \in \Gamma_1(2)$ . Moreover,  $\Gamma_1(2)$  is an embedded codimension-one submanifold and dense in  $\chi_1(2)$ . We still require that any one-parameter family  $X_\lambda$ , ( $\lambda \in (-\varepsilon, \varepsilon)$ ) in  $\chi(1)$  transverse to  $\Gamma_1(2)$  at  $X_0$ , has one of the following normal forms:

- 0.1:  $X_\lambda(x, y) = (1, 0)$  (regular point);
- 0.2:  $X_\lambda(x, y) = (1, x)$  (fold singularity);
- 1.1:  $X_\lambda(x, y) = (1, \lambda + x^2)$  (cusp singularity);
- 1.2:  $X_\lambda(x, y) = (ax, x + by + \lambda)$ ,  $a = \pm 1, b = \pm 2$ ;
- 1.3:  $X_\lambda(x, y) = (x, x - y + \lambda)$ ;
- 1.4:  $X_\lambda(x, y) = (x + y, -x + y + \lambda)$ .

#### Case $n = 2$

**Definition 9** A vector field  $X \in \chi(3)$  belongs to the set  $\Gamma_1(a)$  if the following conditions hold:

- (i)  $X(0) = 0$  and 0 is a hyperbolic critical point of  $X$ ;
- (ii) the eigenvalues of  $DX(0)$  are pairwise distinct and the corresponding eigenspaces are transversal to  $M$  at 0;
- (iii) each pair of non complex conjugate eigenvalues of  $DX(0)$  has distinct real parts.

**Definition 10** A vector field  $X \in \chi(3)$  belongs to the set  $\Gamma_1(b)$  if  $X(0) \neq 0$ ,  $Xh(0) = 0$ ,  $X^2h(0) = 0$  and one of the following conditions hold:

- (1)  $X^3h(0) \neq 0$ ,  $\text{rank}\{Dh(0), DXh(0), DX^2h(0)\} = 2$  and 0 is a non-degenerate critical point of  $Xh|_M$ .
- (2)  $X^3h(0) = 0$ ,  $X^4h(0) \neq 0$  and 0 is a regular point of  $Xh|_M$ .

The next results can be found in [43].

**Theorem 11** The following statements hold:

- (i)  $\Gamma_1(3) = \Gamma_1(a) \cup \Gamma_2(b)$  is a codimension-one submanifold of  $\chi(3)$ .
- (ii)  $\Gamma_1(3)$  is open and dense set in  $\chi_1(3)$  in the topology induced from  $\chi_1(3)$ .
- (iv) For a residual set of smooth curves  $\gamma: \mathbb{R}, 0 \rightarrow \chi(3)$ ,  $\gamma$  meets  $\Gamma_1(3)$  transversally.

Throughout this subsection we fix the function  $h(x, y, z) = z$ .

**Lemma 12 (Classification Lemma)** The elements of  $\Gamma_1(3)$  are classified as follows:

- (a<sub>11</sub>) **Nodal M-Singularity:**  $X(0) = 0$ , the eigenvalues of  $DX(0)$ ,  $\lambda_1, \lambda_2$ , and  $\lambda_3$ , are real, distinct,  $\lambda_1\lambda_j > 0$ ,  $j = 2, 3$  and the eigenspaces are transverse to  $M$  at 0;
- (a<sub>12</sub>) **Saddle M-Singularity:**  $X(0) = 0$ , the eigenvalues of  $DX(0)$ ,  $\lambda_1, \lambda_2$  and  $\lambda_3$ , are real, distinct,  $\lambda_1\lambda_j < 0$ ,  $j = 2$  or  $3$  and the eigenspaces are transverse to  $M$  at 0;
- (a<sub>13</sub>) **Focal M-Singularity:** 0 is a hyperbolic critical point of  $X$ , the eigenvalues of  $DX(0)$  are  $\lambda_{12} = a \pm ib, \lambda_3 = c$ , with  $a, b, c$  distinct from zero and  $c \neq a$ , and the eigenspaces are transverse to  $M$  at 0.
- (b<sub>11</sub>) **Lips M-Singularity:** presented in Definition 8, item 1, when  $\text{Hess}(Fh|_S(0)) > 0$ ;
- (b<sub>12</sub>) **Bec to Bec M-Singularity:** presented in Definition 8, item 1, when  $\text{Hess}(Fh|_S(0)) < 0$ ;
- (b<sub>13</sub>) **Dove's Tail M-Singularity:** presented in Definition 8, item 2.

The next result is proved in [38]. It deals with the normal forms of a codimension-one singularity.

**Theorem 13** i) (Generic Bifurcation and normal forms) Let  $X \in \chi(3)$ . The vector field  $X$  is  $M$ -structurally stable relative to  $\chi_1(3)$  if and only if  $X \in \Gamma_1(3)$ . ii) (Versal unfolding) In the space of one-parameter families of vector fields  $X_\alpha$  in  $\chi(3)$ ,  $\alpha \in (-\varepsilon, \varepsilon)$  an everywhere dense set is formed by generic families such that their normal forms are:

- $X_\alpha \in \Gamma_0(3)$ 
  - 0.1:  $X_\alpha(x, y, z) = (0, 0, 1)$
  - 0.2:  $X_\alpha(x, y, z) = (z, 0, \pm x)$
  - 0.3:  $X_\alpha(x, y, z) = (z, 0, x^2 + y)$

- $X_0 \in \Gamma_1(3)$ 
  - 1.1:  $X_\alpha(x, y, z) = (z, 0, \frac{-3x^2+y^2+\alpha}{2})$
  - 1.2:  $X_\alpha(x, y, z) = (z, 0, \frac{-3x^2-y^2+\alpha}{2})$
  - 1.3:  $X_\alpha(x, y, z) = (z, 0, \frac{4\delta x^3+y+\alpha x}{2})$ , with  $\delta = \pm 1$
  - 1.4:  $X_\alpha(x, y, z) = (axz, byz, \frac{ax+by+cz^2+\alpha}{2})$ , with  $(a, b, c) = \delta(3, 2, 1)$ ,  $\delta = \pm 1$
  - 1.5:  $X_\alpha(x, y, z) = (axz, byz, \frac{ax+by+cz^2+\alpha}{2})$ , with  $(a, b, c) = \delta(1, 3, 2)$ ,  $\delta = \pm 1$
  - 1.6:  $X_\alpha(x, y, z) = (axz, byz, \frac{ax+by+cz^2+\alpha}{2})$ , with  $(a, b, c) = \delta(1, 2, 3)$ ,  $\delta = \pm 1$
  - 1.7:  $X_\alpha(x, y, z) = (xz, 2yz, \frac{x+2y-cz^2+\alpha}{2})$
  - 1.8:  $X_\alpha(x, y, z) = ((-x + y)z, (-x - y)z, \frac{-3x-y+z^2+\alpha}{2})$

### Generic Bifurcation

Let  $Z = (X, Y) \in \Omega^r(n+1)$ . Call by  $\Sigma_0(n+1)$  (resp.  $\Sigma_1(n+1)$ ) the set of all elements that are structurally stable in  $\Omega^r(n+1)$  (resp.  $\Omega_1^r(n+1) = \Omega^r(n+1) \setminus \Sigma_0(n+1)$ ) in  $\Omega^r(n+1)$ . It is clear that a pre-classification of the generic singularities is immediately reached by:

If  $Z = (X, Y) \in \Sigma_0(n+1)$  (resp.  $Z = (X, Y) \in \Sigma_1(n+1)$ ) then  $X$  and  $Y$  are in  $\Gamma_0(n+1)$  (resp.  $X \in \Gamma_0(n+1)$  and  $Y \in \Gamma_1(n+1)$  or vice versa). Of course, the case when both  $X$  and  $Y$  are in  $\Gamma_1(n+1)$  is a codimension-two phenomenon.

### Two-Dimensional Case

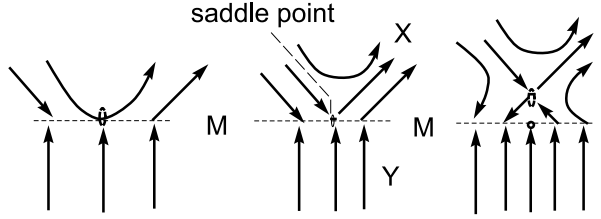
The following result characterizes the structural stability in  $\Omega^r(2)$ .

**Theorem A** (see [31,44]):  $\Sigma_0(2)$  is an open and dense set of  $\Omega^r(2)$ . The vector field  $Z = (X, Y)$  is in  $\Sigma_0(2)$  if and only one of the following conditions is satisfied:

- i) Both elements  $X$  and  $Y$  are regular. When  $0 \in M$  is a simple singularity of  $Z$  then we assume that it is a hyperbolic critical point of  $Z^M$ .
- ii)  $X$  is a fold singularity and  $Y$  is regular (and vice-versa).

The following result still deserves a systematic proof. Following the same strategy stipulated in the generic classification of an  $M$ -singularity, Theorem 11 could be very useful. It is worthwhile to mention [33] where the problem of generic bifurcation in 2D was also addressed.

**Theorem B (Generic Bifurcation)** (see [36,43])  $\Sigma_1(2)$  is an open and dense set of  $\Omega_1^r(2)$ . The vector field  $Z = (X, Y)$  is in  $\Sigma_1(2)$  provided that one of the following conditions is satisfied:



Perturbation Theory for Non-smooth Systems, Figure 4  
M-critical point for  $X$ ,  $M$ -regular for  $Y$  and its unfolding

- i) Both elements  $X$  and  $Y$  are  $M$ -regular. When  $0 \in M$  is a simple singularity of  $Z$  then we assume that it is a codimension-one critical point (saddle-node or a Bogdanov-Takens singularity) of  $Z^M$ .
- ii)  $0$  is a codimension-one  $M$ -singularity of  $X$  and  $Y$  is  $M$ -regular. This case includes when  $0$  is either a cusp  $M$ -singularity or a critical point. Figure 4 illustrates the case when  $0$  is a saddle critical point in the boundary.
- iii) Both  $X$  and  $Y$  are fold  $M$ -singularities at  $0$ . In this case we have to impose that  $0$  is a hyperbolic critical point of the  $C^r$ -extension of  $Z^M$  provided that it is in the boundary of  $M_2 \cup M_3$  (see example below). Moreover when  $0$  is a distinguished fold-fold singularity of  $Z$  then  $0$  is a hyperbolic fixed point of the first return mapping  $\beta_Z$ .

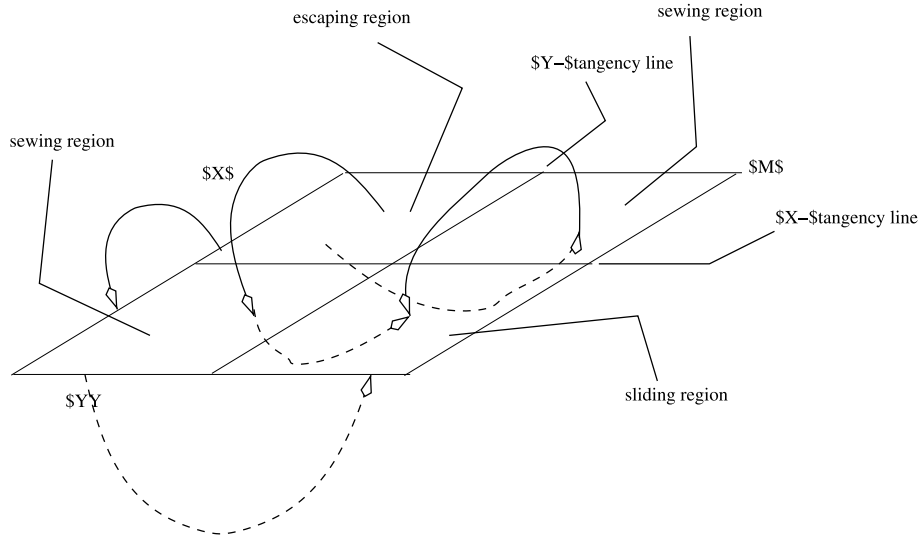
Consider in a small neighborhood of  $0$  in  $\mathbb{R}^2$ , the system  $Z = (X, Y)$  with  $X(x, y) = (1 - x^3 + y^2, x)$ ,  $Y(x, y) = (1 + x + y, -x + x^2)$  and  $h(x, y) = y$ . The point  $0$  is a fold-fold-singularity of  $Z$  with  $M_2 = \{x < 0\}$  and  $Z^s(x, 0) = (2x - x^2)^{-1}(2x - x^4 + x^5)$ . Observe that  $0$  is a hyperbolic critical point of the extended system  $G(x, y) = 2x - x^4 + x^5$ .

The classification of the codimension-two singularities in  $\Omega^r(2)$  is still an open problem. In this direction [51] contains information about the classification of codimension-two  $M$ -singularities.

### Three-Dimensional Case

Let  $Z = (X, Y) \in \Omega^r(3)$ .

The most interesting case to be analyzed is when both vector fields,  $X$  and  $Y$  are fold singularities at  $0$  and the tangency sets  $\tau_X$  and  $\tau_Y$  in  $M$  are in general position at  $0$ . In fact they determine (in  $M$ ) four quadrants, two of them are  $M_1$ -regions, one is an  $M_3$ -region and the other is an  $M_2$ -region (see Fig. 5). We emphasize that the sliding vector field  $Z^M$  can be  $C^r$ -extended to a full neighborhood of  $0$  in  $M$ . Moreover,  $Z^M(0) = 0$ . Inside this class the distinguished fold-fold singularity (as defined in Sub-



**Perturbation Theory for Non-smooth Systems, Figure 5**  
The distinguished fold-fold singularity

sect. “A Construction”) must be taken into account. Denote by  $A$  the set of all *distinguished fold-fold singularities*  $Z \in \Omega^r(3)$ . Moreover, the eigenvalues of  $D\beta_Z(0)$  are  $\lambda = a \pm \sqrt{a^2 - 1}$ . If  $\lambda \in \mathbb{R}$  we say that  $Z$  belongs to  $A_s$ . Otherwise  $Z$  is in  $A_e$ . Recall that  $\beta_Z$  is the first return mapping associated to  $Z$  and  $M$  at 0 as defined in Subsect. “A Construction”.

It is evident that the elements in the open set  $A_e$  are structurally unstable in  $\Omega^r(3)$ . It is worthwhile to mention that in  $A_e$  we detect elements which are asymptotically stable at the origin [48]. Concerning  $A_s$  few things are known.

We have the following result:

**Theorem C** *The vector field  $Z = (X, Y)$  belongs to  $\Sigma_0(3)$  provided that one of the following conditions occurs:*

- i) *Both elements  $X$  and  $Y$  are regular. When  $0 \in M$  is a simple singularity of  $Z$  then we assume that it is a hyperbolic critical point of  $Z^M$ .*
- ii)  *$X$  is a fold singularity at 0 and  $Y$  is regular.*
- iii)  *$X$  is a cusp singularity at 0 and  $Y$  is regular.*
- iv) *Both systems  $X$  and  $Y$  are of fold type at 0. Moreover:*
  - a) *the tangency sets  $\tau_X$  and  $\tau_Y$  are in general position at 0 in  $M$ ;*
  - b) *The eigenspaces associated with  $Z^M$  are transverse to  $\tau_X$  and  $\tau_Y$  at  $0 \in M$  and*
  - c)  *$Z$  is not in  $A$ . Moreover the real parts of non conjugate eigenvalues are distinct.*

We recall that bifurcation diagrams of sliding vector fields are presented in [50,52].

### Singular Perturbation Problem in 2D

Geometric singular perturbation theory is an important tool in the field of continuous dynamical systems. Needless to say that in this area very good surveys are available (refer to [16,18,30]). Here we highlight some results (see [10]) that bridge the space between discontinuous systems in  $\Omega^r(2)$  and singularly perturbed smooth systems.

**Definition 14** Let  $U \subset \mathbb{R}^2$  be an open subset and  $\varepsilon \geq 0$ . A singular perturbation problem in  $U$  (SP-Problem) is a differential system which can be written as

$$x' = \frac{dx}{d\tau} = f(x, y, \varepsilon), \quad y' = \frac{dy}{d\tau} = \varepsilon g(x, y, \varepsilon) \quad (6)$$

or equivalently, after the time re-scaling  $t = \varepsilon \tau$

$$\varepsilon \dot{x} = \varepsilon \frac{dx}{dt} = f(x, y, \varepsilon), \quad \dot{y} = \frac{dy}{dt} = g(x, y, \varepsilon), \quad (7)$$

with  $(x, y) \in U$  and  $f, g$  smooth in all variables.

Our first result is concerned with the transition between non-smooth systems and GSPT.

**Theorem D** *Consider  $Z \in \Omega^r(2)$ ,  $Z_\varepsilon$  its  $\varphi$ -regularization, and  $p \in M$ . Suppose that  $\varphi$  is a polynomial of degree  $k$*

in a small interval  $I \subseteq (-1, 1)$  with  $0 \in I$ . Then the trajectories of  $Z_\varepsilon$  in  $V_\varepsilon = \{q \in \mathbb{R}^2, 0: h(q)/\varepsilon \in I\}$  are in correspondence with the solutions of an ordinary differential equation  $z' = \alpha(z, \varepsilon)$ , satisfying that  $\alpha$  is smooth in both variables and  $\alpha(z, 0) = 0$  for any  $z \in M$ . Moreover, if  $((X - Y)h^k)(p) \neq 0$  then we can take a  $C^{r-1}$ -local coordinate system  $\{(\partial/\partial x)(p), (\partial/\partial y)(p)\}$  such that this smooth ordinary differential equation is a SP-problem.

The understanding of the phase portrait of the vector field associated to a SP-problem is the main goal of the *geometric singular perturbation-theory* (GSP-theory). The techniques of GSP-theory can be used to obtain information on the dynamics of (6) for small values of  $\varepsilon > 0$ , mainly in searching minimal sets.

System (6) is called the *fast system*, and (7) the *slow system* of the SP-problem. Observe that for  $\varepsilon > 0$  the phase portraits of the fast and the slow systems coincide.

Theorem D says that we can transform a discontinuous vector field in a SP-problem. In general this transition cannot be done explicitly. Theorem E provides an explicit formula of the SP-problem for a suitable class of vector fields. Before the statement of such a result we need to present some preliminaries.

Consider  $C = \{\xi: \mathbb{R}^2, 0 \rightarrow \mathbb{R}\}$  with  $\xi \in C^r$  and  $L(\xi) = 0$  where  $L(\xi)$  denotes the linear part of  $\xi$  at  $(0, 0)$ .

Let  $\Omega_d \subset \Omega^r(2)$  be the set of vector fields  $Z = (X, Y)$  in  $\Omega^r(2)$  such that there exists  $\xi \in C$  that is a solution of

$$\nabla \xi(X - Y) = \Pi_i(X - Y), \quad (8)$$

where  $\nabla \xi$  is the gradient of the function and  $\Pi_i$  denote the canonical projections, for  $i = 1$  or  $i = 2$ .

**Theorem E** Consider  $Z \in \Omega_d$  and  $Z_\varepsilon$  its  $\varphi$ -regularization. Suppose that  $\varphi$  is a polynomial of degree  $k$  in a small interval  $I \subset \mathbb{R}$  with  $0 \in I$ . Then the trajectories of  $Z_\varepsilon$  on  $V_\varepsilon = \{q \in \mathbb{R}^2, 0: h(q)/\varepsilon \in I\}$  are solutions of a SP-problem.

We remark that the singular problems discussed in the previous theorems, when  $\varepsilon \searrow 0$ , defines a dynamical system on the discontinuous set of the original problem. This fact can be very useful for problems in Control Theory.

Our third theorem says how the fast and the slow systems approximate the discontinuous vector field. Moreover, we can deduce from the proof that whereas the fast system approximates the discontinuous vector field, the slow system approaches the corresponding sliding vector field.

Consider  $Z \in \Omega^r(2)$  and  $\rho: \mathbb{R}^2, 0 \rightarrow \mathbb{R}$  with  $\rho(x, y)$  being the distance between  $(x, y)$  and  $M$ . We denote by  $\widehat{Z}$  the vector field given by  $\widehat{Z}(x, y) = \rho(x, y)Z(x, y)$ .

In what follows we identify  $\widehat{Z}_\varepsilon$  and the vector field on  $\{\mathbb{R}^2, 0\} \setminus M \times \mathbb{R} \subset \mathbb{R}^3$  given by  $\widehat{Z}(x, y, \varepsilon) = (\widehat{Z}_\varepsilon(x, y), 0)$ .

**Theorem F** Consider  $p = 0 \in M$ . Then there exists an open set  $U \subset \mathbb{R}^2$ ,  $p \in U$ , a three-dimensional manifold  $M$ , a smooth function  $\Phi: U \rightarrow \mathbb{R}^3$  and a SP-problem  $W$  on  $M$  such that  $\Phi$  sends orbits of  $W|_{\Phi^{-1}(U \times (0, +\infty))}$  in orbits of  $\widehat{Z}|_{(U \times (0, +\infty))}$ .

### Examples

1. Take  $X(x, y) = (1, x)$ ,  $Y(x, y) = (-1, -3x)$ , and  $h(x, y) = y$ . The discontinuity set is  $\{(x, 0) \mid x \in \mathbb{R}\}$ . We have  $Xh = x$ ,  $Yh = -3x$ , and then the unique non-regular point is  $(0, 0)$ . In this case we may apply Theorem E.
2. Let  $Z_\varepsilon(x, y) = (y/\varepsilon, 2xy/\varepsilon - x)$ . The associated partial differential equation (refer to Theorem E) with  $i = 2$  given above becomes  $2(\partial \xi / \partial x) + 4x(\partial \xi / \partial y) = 4x$ . We take the coordinate change  $\bar{x} = x$ ,  $\bar{y} = y - x^2$ . The trajectories of  $X_\varepsilon$  in these coordinates are the solutions of the singular system

$$\varepsilon \dot{\bar{x}} = \bar{y} + \bar{x}^2, \quad \dot{\bar{y}} = -\bar{x}.$$

3. In what follows we try, by means of an example, to present a rough idea on the transition from non-smooth systems to GSPT. Consider  $X(x, y) = (3y^2 - y - 2, 1)$ ,  $Y(x, y) = (-3y^2 - y + 2, -1)$  and  $h(x, y) = x$ . The regularized vector field is

$$Z_\varepsilon(x, y) = \left( \frac{1}{2} + \frac{1}{2} \varphi \left( \frac{x}{\varepsilon} \right) \right) (3y^2 - y - 2, 1) + \left( \frac{1}{2} - \frac{1}{2} \varphi \left( \frac{x}{\varepsilon} \right) \right) (-3y^2 - y + 2, -1).$$

After performing the polar blow up coordinates  $\alpha: [0, +\infty) \times [0, \pi] \times \mathbb{R} \rightarrow \mathbb{R}^3$  given by  $x = r \cos \theta$  and  $\varepsilon = r \sin \theta$  the last system is expressed by:

$$r\dot{\theta} = -\sin \theta(-y + \varphi(\cot \theta)(3y^2 - 2)), \quad \dot{y} = \varphi(\cot \theta).$$

So the slow manifold is given implicitly by  $\varphi(\cot \theta) = y/(3y^2 - 2)$  which defines two functions  $y_1(\theta) = (1 + \sqrt{1 + 24\varphi^2(\cot \theta)})/(6\varphi(\cot \theta))$  and  $y_2(\theta) = (1 - \sqrt{1 + 24\varphi^2(\cot \theta)})/(6\varphi(\cot \theta))$ . The function  $y_1(\theta)$  is increasing,  $y_1(0) = 1$ ,  $\lim_{\theta \rightarrow \pi/2^-} y_1(\theta) = +\infty$ ,  $\lim_{\theta \rightarrow \pi/2^+} y_1(\theta) = -\infty$  and  $y_1(\pi) = -1$ . The function

$y_2(\theta)$  is increasing,  $y_2(0) = -2/3$ ,  $\lim_{\theta \rightarrow \pi/2} y_2(\theta) = 0$  and  $y_2(\pi) = 2/3$ . We can extend  $y_2$  to  $(0, \pi)$  as a differential function with  $y_2(\pi/2) = 0$ .

The fast vector field is  $(\theta', 0)$  with  $\theta' > 0$  if  $(\theta, y)$  belongs to

$$\left[ \left(0, \frac{\pi}{2}\right) \times (y_2(\theta), y_1(\theta)) \cup \left(\frac{\pi}{2}, \pi\right) \times (y_2(\theta), +\infty) \right. \\ \left. \cup \left(\frac{\pi}{2}, \pi\right) \times (-\infty, y_1(\theta)) \right]$$

and with  $\theta' < 0$  if  $(\theta, y)$  belongs to

$$\left[ \left(0, \frac{\pi}{2}\right) \times (y_1(\theta), +\infty) \cup \left(0, \frac{\pi}{2}\right) \times (-\infty, y_2(\theta)) \right. \\ \left. \cup \left(\frac{\pi}{2}, \pi\right) \times (y_1(\theta), y_2(\theta)) \right].$$

The reduced flow has one singular point at  $(0, 0)$  and it takes the positive direction of the  $y$ -axis if  $y \in (-\frac{2}{3}, 0) \cup (1, \infty)$  and the negative direction of the  $y$ -axis if  $y \in (-\infty, -1) \cup (0, \frac{2}{3})$ .

One can see that the singularities  $(\theta, y, r) = (0, 1, 0)$  and  $(\theta, y, r) = (0, -1, 0)$  are not normally hyperbolic points. In this way, as usual, we perform additional blow ups. In Fig. 6 we illustrate the fast and the slow dynamics of the SP-problem. We present a phase portrait on the blowing up locus where a double arrow over a trajectory means that the trajectory belongs to the fast dynamical system, and a simple arrow means that the trajectory belongs to the slow dynamical system.

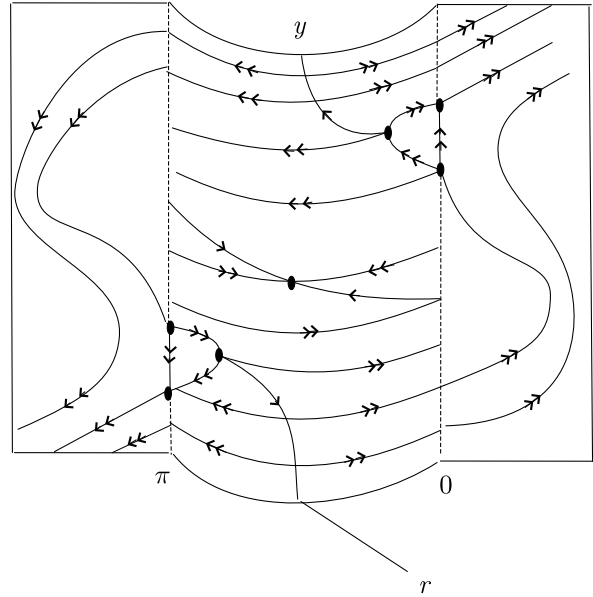
### Future Directions

Our concluding section is devoted to an outlook. Firstly we present some open problems linked with the setting that point out future directions of research. The main task for the future seems to bring the theory of non-smooth dynamical systems to a similar maturity as that of smooth systems. Finally we briefly discuss the main results in this text.

### Some Problems

In connection to this present work, some theoretical problems remain open:

1. The description of the bifurcation diagram of the codimension-two singularities in  $\Omega(2)$ . In this last class we find some models (see [37]) where the following questions can also be addressed. a) When is a typical sin-



**Perturbation Theory for Non-smooth Systems, Figure 6**  
Example of fast and slow dynamics of the SP-Problem

gularity topologically equivalent to a regular center? b) How about the isochronicity of such a center? c) When does a polynomial perturbation of such a system in  $\Omega(2)$  produce limit cycles? The articles [9,13,21,22,47] can be useful auxiliary references.

2. Let  $\Omega(N)$  be the set of all non-smooth vector fields on a two-dimensional compact manifold  $N$  having a codimension-one compact submanifold  $M$  as its discontinuity set. The problem is to study the global generic bifurcation in  $\Omega(N)$ . The articles [31,33,40,46] can be useful auxiliary references.
3. Study of the bifurcation set in  $\Omega^r(3)$ . The articles [38,40,43,50] can be useful auxiliary references.
4. Study of the dynamics of the distinguished *fold-fold* singularity in  $\Omega^r(n+1)$ . The article [48] can be a useful auxiliary reference.
5. In many applications examples of non-smooth systems where the discontinuities are located on algebraic varieties are available. For instance, consider the system  $\ddot{x} + x \text{sign}(x) + \text{sign}(\dot{x}) = 0$ . Motivated by such models we present the following problem. Let 0 be a non-degenerate critical point of a smooth mapping  $h: \mathbb{R}^{n+1}, 0 \rightarrow \mathbb{R}, 0$ . Let  $\Phi(n+1)$  be the space of all vector fields  $Z$  on  $\mathbb{R}^{n+1}, 0$  defined in the same way as  $\Omega(n+1)$ . We propose the following. i) Classify the typical singularities in that space. ii) Analyze the elements of  $\Phi(2)$  by means of “regularization processes” and the methods of GSPT, similarly to Sect. “Vector Fields near



the Boundary". The articles [1,2] can be very useful auxiliary references.

6. In [27,29] classes of 4D-relay systems are considered. Conditions for the existence of one-parameter families of periodic orbits terminating at typical singularities are provided. We propose to find conditions for the existence of such families for  $n$ -dimensional relay systems.

## Conclusion

In this paper we have presented a compact survey of the geometric/qualitative theoretical features of non-smooth dynamical systems. We feel that our survey illustrates that this field is still in its early stages but enjoying growing interest. Given the importance and the relevance of such a theme, we have pointed above some open questions and we remark that there is still a wide range of bifurcation problems to be tackled. A brief summary of the main results in the text is given below.

1. We firstly deal with two-dimensional non-smooth vector fields  $Z = (X, Y)$  defined around the origin in  $R^2$ , where the discontinuity set is concentrated on the line  $\{y = 0\}$ . The first task is to characterize those systems which are structurally stable. This characterization is a starting point with which to establish a bifurcation theory as indicated by the Thom–Smale program.
2. In higher dimension the problem becomes much more complicated. We have presented here sufficient conditions for the three-dimensional local structural stability. Any further investigation on bifurcation in this context must pass through a deep analysis of the so called *fold-fold* singularity.
3. We have established a bridge between discontinuous and singularly perturbed smooth systems. Many similarities between such systems were observed and a comparative study of the two categories is called for.

## Bibliography

### Primary Literature

1. Alexander JC, Seidman TI (1998) Sliding modes in intersecting switching surfaces I: Blending. *Houston J Math* 24(3):545–569
2. Alexander JC, Seidman TI (1999) Sliding modes in intersecting switching surfaces II: Hysteresis. *Houston J Math* 25:185–211
3. Andronov A, Pontryagin S (1937) Structurally stable systems. *Dokl Akad Nauk SSSR* 14:247–250
4. Andronov AA, Vitt AA, Khaikin SE (1966) *Theory of oscillators*, Dover, New York
5. Anosov DV (1959) Stability of the equilibrium positions in relay systems. *Autom Remote Control* XX(2):135–149
6. Arnold VI (1983) *Methods in the theory of ordinary differential equations*. Springer, New York
7. Bazykin AD (1998) *Nonlinear dynamics of interacting populations*. World Sc. Publ. Co. Inc., River-Edge, NJ
8. Bonnard B, Chyba M (2000) *Singular trajectories and their role in control theory*. Mathématiques and Applications, vol 40. Springer, Berlin
9. Broucke ME, Pugh CC, Simić SN (2001) Structural stability of piecewise smooth systems. *Comput Appl Math* 20(1–2):51–89
10. Buzzi C, Silva PR, Teixeira MA (2006) A singular approach to discontinuous vector fields on the plane. *J Differ Equ* 23:633–655
11. Chillingworth DR (2002) J.(4-SHMP) Discontinuity geometry for an impact oscillator. *Dyn Syst* 17(4):389–420
12. Chua LO (1992) The genesis of Chua's circuit. *AEU* 46:250
13. Coll B, Gasull A, Prohens R (2001) Degenerate Hopf bifurcations in discontinuous planar systems. *J Math Anal Appl* 253(2):671–690
14. di Bernardo M, Budd C, Champneys AR, Kowalczyk P, Nordmark AB, Olivar G, Piiroinen PT (2005) *Bifurcations in non-smooth dynamical systems*. Bristol Centre for Applied Nonlinear Mathematics, N. 2005-4, Bristol
15. Dumortier F (1977) Singularities of vector fields. *IMPA, Rio de Janeiro*
16. Dumortier F, Roussarie R (1996) Canard cycles and center manifolds. *Mem Amer Mat Soc* 121:x+100
17. Ekeland I (1977) Discontinuités de champs hamiltoniens et existence de solutions optimales en calcul des variations. *Inst Hautes Études Sci Publ Math* 47:5–32
18. Fenichel N (1979) Geometric singular perturbation theory for ordinary differential equations. *J Differ Equ* 31:53–98
19. Filippov AF (1988) *Differential equations with discontinuous righthand sides*. Kluwer Academic, Dordrecht
20. Flügge-Lotz I (1953) *Discontinuous automatic control*. Princeton University, Princeton, pp vii+168
21. Garcia R, Teixeira MA (2004) Vector fields in manifolds with boundary and reversibility-an expository account. *Qual Theory Dyn Syst* 4:311–327
22. Gasull A, Torregrosa J (2003) Center-focus problem for discontinuous planar differential equations. *Int J Bifurc Chaos Appl Sci Eng* 13(7):1755–1766
23. Henry P (1972) Diff. equations with discontinuous right-hand side for planning procedures. *J Econ Theory* 4:545–551
24. Hogan S (1989) On the dynamics of rigid-block motion under harmonic forcing. *Proc Roy Soc London A* 425:441–476
25. Ito T (1979) A Filippov solution of a system of diff. eq. with discontinuous right-hand sides. *Econ Lett* 4:349–354
26. Jackson JD (1999) *Classical electrodynamics*, 3rd edn. Wiley, New York, pp xxii+808
27. Jacquemard A, Teixeira MA (2003) Computer analysis of periodic orbits of discontinuous vector fields. *J Symbol Comput* 35:617–636
28. Jacquemard A, Teixeira MA (2003) On singularities of discontinuous vector fields. *Bull Sci Math* 127:611–633
29. Jacquemard A, Teixeira MA (2005) Invariant varieties of discontinuous vector fields. *Nonlinearity* 18:21–43
30. Jones C (1995) *Geometric singular perturbation theory*. C.I.M.E. Lectures, Montecatini Terme, June 1994, Lecture Notes in Mathematics 1609. Springer, Heidelberg

31. Kozlova VS (1984) Roughness of a discontinuous system. *Vestnik Moskovskogo Universiteta, Matematika* 5:16–20
32. Kunze M, Kupper T (1997) Qualitative bifurcation analysis of a non-smooth friction oscillator model. *Z Angew Math Phys* 48:87–101
33. Kuznetsov YA et al (2003) One-parameter bifurcations in planar Filippov systems. *Int J Bifurc Chaos* 13:2157–2188
34. Llibre J, Teixeira MA (1997) Regularization of discontinuous vector fields in dimension 3. *Discret Contin Dyn Syst* 3(2):235–241
35. Luo CJ (2006) Singularity and dynamics on discontinuous vector fields. *Monograph Series on Nonlinear Science and Complexity*. Elsevier, New York, pp i+310
36. Machado AL, Sotomayor J (2002) Structurally stable discontinuous vector fields in the plane. *Qual Theory Dyn Syst* 3:227–250
37. Manosas F, Torres PJ (2005) Isochronicity of a class of piecewise continuous oscillators. *Proc of AMS* 133(10):3027–3035
38. Medrado J, Teixeira MA (1998) Symmetric singularities of reversible vector fields in dimension three. *Physica D* 112:122–131
39. Medrado J, Teixeira MA (2001) Codimension-two singularities of reversible vector fields in 3D. *Qualit Theory Dyn Syst* J 2(2):399–428
40. Percell PB (1973) Structural stability on manifolds with boundary. *Topology* 12:123–144
41. Seidman T (2006) Aspects of modeling with discontinuities. In: N'Guerekata G (ed) *Advances in Applied and Computational Mathematics Proc Dover Conf*, Cambridge. <http://www.umbc.edu/~seideman>
42. Sotomayor J (1974) Structural stability in manifolds with boundary. In: *Global analysis and its applications*, vol III. IEAA, Vienna, pp 167–176
43. Sotomayor J, Teixeira MA (1988) Vector fields near the boundary of a 3-manifold. *Lect Notes in Math*, vol 331. Springer, Berlin-Heidelberg, pp 169–195
44. Sotomayor J, Teixeira MA (1996) Regularization of discontinuous vector fields. *International Conference on Differential Equations*, Lisboa. World Sc, Singapore, pp 207–223
45. Sussmann H (1979) Subanalytic sets and feedback control. *J Differ Equ* 31:31–52
46. Teixeira MA (1977) Generic bifurcations in manifolds with boundary. *J Differ Equ* 25:65–89
47. Teixeira MA (1979) Generic bifurcation of certain singularities. *Boll Unione Mat Ital* 16-B:238–254
48. Teixeira MA (1990) Stability conditions for discontinuous vector fields. *J Differ Equ* 88:15–24
49. Teixeira MA (1991) Generic Singularities of Discontinuous Vector Fields. *An Ac Bras Cienc* 53(2):257–260
50. Teixeira MA (1993) Generic bifurcation of sliding vector fields. *J Math Anal Appl* 176:436–457
51. Teixeira MA (1997) Singularities of reversible vector fields. *Physica D* 100:101–118
52. Teixeira MA (1999) Codimension-two singularities of sliding vector fields. *Bull Belg Math Soc* 6(3):369–381
53. Vishik SM (1972) Vector fields near the boundary of a manifold. *Vestnik Moskovskogo Universiteta, Matematika* 27(1):13–19
54. Zelikin MI, Borisov VF (1994) Theory of chattering control with applications to astronautics, robotics, economics, and engineering. Birkhäuser, Boston

## Books and Reviews

- Agrachev AA, Sachkov YL (2004) Control theory from the geometric viewpoint. *Encyclopaedia of Mathematical Sciences*, 87. Control theory and optimization, vol II. Springer, Berlin, pp xiv+412
- Barbashin EA (1970) Introduction to the theory of stability. Wolters-Noordhoff Publishing, Groningen. Translated from the Russian by Transcripta Service, London. Edited by T. Lukes, pp 223
- Brogliato B (ed) (2000) Impacts in mechanical systems. *Lect Notes in Phys*, vol 551. Springer, Berlin, pp 160
- Carmona V, Freire E, Ponce E, Torres F (2005) Bifurcation of invariant cones in piecewise linear homogeneous systems. *Int J Bifurc Chaos* 15(8):2469–2484
- Davydov AA (1994) Qualitative theory of control systems. American Mathematical Society, Providence, RI. (English summary) Translated from the Russian manuscript by V. M. Volosov. *Translations of Mathematical Monographs*, 141, pp viii+147
- Dercole F, Gragnani F, Kuznetsov YA, Rinaldi S (2003) Numerical sliding bifurcation analysis: an application to a relay control system. *IEEE Trans Circuit Systems – I: Fund Theory Appl* 50:1058–1063
- Glocker C (2001) Set-valued force laws: dynamics of non-smooth systems. *Lecture Notes in Applied Mechanics*, vol 1. Springer, Berlin
- Hogan J, Champneys A, Krauskopf B, di Bernardo M, Wilson E, Osinga H, Homer M (eds) *Nonlinear dynamics and chaos: Where do we go from here?* Institute of Physics Publishing (IOP), Bristol, pp xi+358, 2003
- Huertas JL, Chen WK, Madan RN (eds) (1997) Visions of nonlinear science in the 21st century, Part I. *Festschrift dedicated to Leon O. Chua on the occasion of his 60th birthday*. Papers from the workshop held in Sevilla, June 26, 1996. *Int J Bifurc Chaos Appl Sci Eng* 7 (1997), no. 9. World Scientific Publishing Co. Pte. Ltd., Singapore, pp i–iv, pp 1907–2173
- Komuro M (1990) Periodic bifurcation of continuous piece-wise vector fields. In: Shiraiwa K (ed) *Advanced series in dynamical systems*, vol 9. World Sc, Singapore
- Kunze M (2000) Non-smooth dynamical systems. *Lecture Notes in Mathematics*, vol 1744. Springer, Berlin, pp x+228
- Kunze M, Küpper T (2001) Non-smooth dynamical systems: an overview. In: Fiedler B (ed) *Ergodic theory, analysis, and efficient simulation of dynamical systems*. Springer, Berlin, pp 431–452
- Llibre J, Silva PR, Teixeira MA (2007) Regularization of discontinuous vector fields on  $R^3$  via singular perturbation. *J Dyn Differ Equ* 19(2):309–331
- Martinet J (1974) Singularités des fonctions et applications différentiables. *Pontificia Universidade Católica do Rio de Janeiro*, Rio de Janeiro, pp xiii+181 (French)
- Minorsky N (1962) *Nonlinear oscillations*. D Van Nostrand Co. Inc., Princeton, N.J., Toronto, London, New York, pp xviii+714
- Minorsky N (1967) *Théorie des oscillations*. *Mémorial des Sciences Mathématiques Fasc*, vol 163. Gauthier-Villars, Paris, pp i+114 (French)
- Minorsky N (1969) *Theory of nonlinear control systems*. McGraw-Hill, New York, London, Sydney, pp xx+331
- Ostrowski JP, Burdick JW (1997) Controllability for mechanical systems with symmetries and constraints. *Recent developments in robotics. Appl Math Comput Sci* 7(2):305–331

- Peixoto MC, Peixoto MM (1959) Structural Stability in the plane with enlarged conditions. *Anais da Acad Bras Ciências* 31:135–160
- Rega G, Lenci S (2003) Nonsmooth dynamics, bifurcation and control in an impact system. *Syst Anal Model Simul Arch* 43(3), Gordon and Breach Science Publishers, Inc. Newark, pp 343–360
- Seidman T (1995) Some limit problems for relays. In: Lakshmikantham V (ed) *Proc First World Congress of Nonlinear Analysis*, vol I. Walter de Gruyter, Berlin, pp 787–796
- Sotomayor J (1974) Generic one-parameter families of vector fields on 2-manifolds. *Publ Math IHES* 43:5–46
- Szmolyan P (1991) Transversal heteroclinic and homoclinic orbits in singular perturbation problems. *J Differ Equ* 92:252–281
- Teixeira MA (1985) Topological invariant for discontinuous vector fields. *Nonlinear Anal TMA* 9(10):1073–1080
- Utkin V (1978) Sliding modes and their application in variable structure systems. Mir, Moscow
- Utkin V (1992) *Sliding Modes in Control and Optimization*. Springer, Berlin

## Perturbation Theory for PDEs

DARIO BAMBUSI

Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italia

### Article Outline

Glossary

Definition of the Subject

Introduction

The Hamiltonian Formalism for PDEs

Normal Form

for Finite Dimensional Hamiltonian Systems

Normal Form for Hamiltonian PDEs: General Comments

Normal Form for Resonant Hamiltonian PDEs and its Consequences

Normal Form for Nonresonant Hamiltonian PDEs

Non Hamiltonian PDEs

Extensions and Related Results

Future Directions

Bibliography

### Glossary

**Perturbation theory** The study of a dynamical systems which is a perturbation of a system whose dynamics is known. Typically the unperturbed system is linear or integrable.

**Normal form** The normal form method consists of constructing a coordinate transformation which changes the equations of a dynamical system into new equations which are as simple as possible. In Hamiltonian

systems the theory is particularly effective and typically leads to a very precise description of the dynamics.

**Hamiltonian PDE** A Hamiltonian PDE is a partial differential equation (abbreviated PDE) which is equivalent to the Hamilton equation of a suitable Hamiltonian function. Classical examples are the nonlinear wave equation, the Nonlinear Schrödinger equation, and the Kortweg–de Vries equation.

### Resonance vs. Non-Resonance

A frequency vector  $\{\omega_k\}_{k=1}^n$  is said to be non-resonant if its components are independent over the relative integers. On the contrary, if there exists a non-vanishing  $K \in \mathbb{Z}^n$  such that  $\omega \cdot K = 0$  the frequency vector is said to be resonant. Such a property plays a fundamental role in normal form theory. Non-resonance typically implies stability.

**Actions** The action of a harmonic oscillator is its energy divided by its frequency. It is usually denoted by  $I$ . The typical issue of normal form theory is that in nonresonant systems the actions remain approximatively unchanged for very long times. In resonant systems there are linear combinations of the actions with such properties.

**Sobolev space** Space of functions which have weak derivatives enjoying suitable integrability properties. Here we will use the spaces  $H^s$ ,  $s \in \mathbb{N}$  of the functions which are square integrable together with their first  $s$  weak derivatives.

### Definition of the Subject

Perturbation theory for PDEs is a part of the qualitative theory of differential equations. One of the most effective methods of perturbation theory is the normal form theory which consists of using coordinate transformations in order to describe the qualitative features of a given or generic equation. Classical normal form theory for ordinary differential equations has been used all along the last century in many different domains, leading to important results in pure mathematics, celestial mechanics, plasma physics, biology, solid state physics, chemistry and many other fields.

The development of effective methods to understand the dynamics of partial differential equations is relevant in pure mathematics as well as in all the fields in which partial differential equations play an important role. Fluidodynamics, oceanography, meteorology, quantum mechanics, and electromagnetic theory are just a few examples of potential applications. More precisely, the normal form theory allows one to understand whether a small nonlinearity can change the dynamics of a linear PDE or not.

Moreover, it allows one to understand how the changes can be avoided or forced. Finally, when the changes are possible it allows to predict the behavior of the perturbed system.

## Introduction

The normal form method was developed by Poincaré and Birkhoff between the end of the 19th century and the beginning of the 20th century. During the last 20 years the method has been successfully generalized to a suitable class of partial differential equations (PDEs) in finite volume (in the case of infinite volume dispersive effects appear and the theory is very different. See e. g. [58]). In this article we will give an introduction to this recent field. We will almost only deal with Hamiltonian PDEs, since on the one hand the theory for non Hamiltonian systems is a small variant of the one we will present here, and on the other hand most models are Hamiltonian.

We will start by a generalization of the Hamiltonian formalism to PDEs, followed by a review of the classical theory and by the actual generalization of normal form theory to PDEs.

In the next section we give a generalization of the Hamiltonian formalism to PDEs. The main new fact is that in PDEs the Hamiltonian is usually a smooth function, but the corresponding vector field is nonsmooth (it is an operator extracting derivatives). So the standard formalism has to be slightly modified [10,29,45,49,52,61]. Here we will present a version of the Hamiltonian formalism which is enough to cover the models of interest for local perturbation theory. To clearly illustrate the situation we will start the article with an introduction to the Lagrangian and Hamiltonian formalism for the wave equation. This will lead to the introduction of the paradigm Hamiltonian which is usually studied in this context. This will be followed by a few results on the Hamiltonian formalism that are needed for perturbation theory.

Subsequently, we shortly present the standard Birkhoff normal form theory for finite dimensional systems. This is useful since all the formal aspects are equal in the classical case and the case of PDEs.

Then we come to the generalization of normal form theory to PDEs. In the present paper we will concentrate almost only on the case of 1-dimensional semilinear equations. This is due to the fact that the theory of higher dimensional and quasilinear equations is still quite unsatisfactory.

In PDEs one essentially meets two kinds of difficulties. The first one is related to the existence of non smooth vec-

tor fields. The second difficulty is due to the fact that in the infinite dimensional case there are small denominators which are much worse than in the finite dimensional one.

We first present the theory for completely resonant systems [10,14] in which the difficulties related to small denominators do not appear. It turns out that it is quite easy to obtain a normal form theorem for resonant PDEs, but the kind of normal form one gets is usually quite poor. In order to extract dynamical informations from the normal form one can only compute and study it explicitly. Usually this is very difficult. Nevertheless in some cases it is possible and leads to quite strong results. We will illustrate such a situation by studying a nonlinear Schrödinger equation [2,25].

For the general case there is a theorem ensuring that a generic system admits at least one family of “periodic like trajectories” which are stable over exponentially long times [15]. We will give its statement and an application to the nonlinear wave equation

$$u_{tt} - u_{xx} + \mu^2 u + f(u) = 0, \quad (1)$$

with  $\mu = 0$  and the Dirichlet boundary conditions on a segment [55].

Then we turn to the case of nonresonant PDEs. The main difficulty is that small denominators accumulate to zero already at order 3. Such a problem has been overcome in [4,6,9,12,43] by taking advantage of the fact that the nonlinearities appearing in PDEs typically have a special form. In this case one can deduce a very precise description of the dynamics and also some interesting results of the kind of almost global existence of smooth solutions [46]. To illustrate the theory we will make reference to the nonlinear wave Eq. (1) with almost any  $\mu$ , and to the nonlinear Schrödinger equation.

Another aspect of the theory of close to integrable Hamiltonian PDEs concerns the extension of KAM theory to PDEs. We will not present it here. We just recall the most celebrated results which are those due to Kuksin [48], Wayne [60], Craig–Wayne [34], Bourgain [24,26], Kuksin–Pöschel [50], Eliasson–Kuksin [39]. All these results ensure the existence of families of quasiperiodic solutions, i. e. solutions lying on finite dimensional manifolds. We also mention the papers [21,57] where some Cantor families of full dimension tori are constructed. We point out that in the dynamics on such  $\infty$ -dimensional tori the amplitude of oscillation of the linear modes decreases super exponentially with their index. A remarkable exception is provided by the paper [27] where the tori constructed are more “thick” (even if of course they lie on Cantor families).

On the contrary, the results of normal form theory describe solutions starting on opens subsets of the phase space, and do not have particularly strong localizations properties with respect to the index. The price one has to pay is that the description one gets turns out to be valid only over long but finite times.

Finally we point out a related research stream that has been carried on by Bourgain [21,22,23,25] who studied intensively the behavior of high Sobolev norms in close to integrable Hamiltonian PDEs (see also [28]).

## The Hamiltonian Formalism for PDEs

### The Gradient of a Functional

**Definition 1** Consider a function  $f \in C^\infty(\mathcal{U}_s, \mathbb{R})$ ,  $\mathcal{U}_s \subset H^s(\mathbb{T})$  open,  $s \geq 0$  a fixed parameter and  $\mathbb{T} := \mathbb{R}/2\pi\mathbb{Z}$  is the 1-dimensional torus. We will denote by  $\nabla f(u)$  the gradient of  $f$  with respect to the  $L^2$  metric, namely the unique function such that

$$\langle \nabla f(u), h \rangle_{L^2} = df(u)h, \quad \forall h \in H^s \quad (2)$$

where

$$\langle u, v \rangle_{L^2} := \int_{-\pi}^{\pi} u(x)v(x)dx \quad (3)$$

is the  $L^2$  scalar product and  $df(u)$  is the differential of  $f$  at  $u$ . The gradient is a smooth map from  $H^s$  to  $H^{-s}$  (see e.g. [3]).

**Example 2** Consider the function

$$f(u) := \int_{-\pi}^{\pi} \frac{u_x^2}{2} dx, \quad (4)$$

which is differentiable as a function from  $H^s \rightarrow \mathbb{R}$  for any  $s \geq 1$ . One has

$$df(u)h = \int_{-\pi}^{\pi} u_x h_x dx = \int_{-\pi}^{\pi} -u_{xx} h dx = \langle -u_{xx}, h \rangle_{L^2} \quad (5)$$

and therefore in this case one has  $\nabla f(u) = -u_{xx}$ .

**Example 3** Let  $\mathcal{F}: \mathbb{R}^2 \rightarrow \mathbb{R}$  be a smooth function and define

$$f(u) = \int_{-\pi}^{\pi} \mathcal{F}(u, u_x) dx \quad (6)$$

then the gradient of  $f$  coincides with the so called functional derivative of  $\mathcal{F}$ :

$$\nabla f \equiv \frac{\delta \mathcal{F}}{\delta u} := \frac{\partial \mathcal{F}}{\partial u} - \frac{\partial}{\partial x} \frac{\partial \mathcal{F}}{\partial u_x}. \quad (7)$$

### Lagrangian and Hamiltonian Formalism for the Wave Equation

Until Subsect. “Basic Elements of Hamiltonian Formalism for PDEs” we will work at a formal level, without specifying the function spaces and the domains.

**Definition 4** Let  $L(u, \dot{u})$  be a Lagrangian function, then the corresponding Lagrange equations are given

$$\nabla_u L - \frac{d}{dt} \nabla_{\dot{u}} L = 0 \quad (8)$$

where  $\nabla_u L$  is the gradient with respect to  $u$  only, and similarly  $\nabla_{\dot{u}}$  is the gradient with respect to  $\dot{u}$ .

**Example 5** Consider the Lagrangian

$$L(u, \dot{u}) := \int_{-\pi}^{\pi} \left( \frac{\dot{u}^2}{2} - \frac{u_x^2}{2} - \mu^2 \frac{u^2}{2} - F(u) \right) dx. \quad (9)$$

then the corresponding Lagrange equations are given by (1) with  $f = -F'$ .

Given a Lagrangian system with a Lagrangian function  $L$ , one defines the corresponding Hamiltonian system as follows.

**Definition 6** Consider the momentum  $v := \nabla_{\dot{u}} L$  conjugated to  $u$ ; assume that  $L$  is convex with respect to  $\dot{u}$ , then the Hamiltonian function associated to  $L$  is defined by

$$H(v, u) := [\langle v; \dot{u} \rangle_{L^2} - L(u, \dot{u})]_{\dot{u}=\dot{u}(u,v)}. \quad (10)$$

**Definition 7** Let  $H(v, u)$  be a Hamiltonian function, then the corresponding Hamilton equations are given by

$$\dot{v} = -\nabla_u H, \quad \dot{u} = \nabla_v H. \quad (11)$$

As in the finite dimensional case, one has that the Lagrange equations are equivalent to the Hamilton equation of  $H$ .

An elementary computation shows that for the wave equation one has  $v = \dot{u}$  and

$$H(v, u) = \int_{-\pi}^{\pi} \left( \frac{v^2 + u_x^2 + \mu^2 u^2}{2} + F(u) \right) dx \quad (12)$$



### Canonical Coordinates

Consider a Lagrangian system and let  $\mathbf{e}_k$  be an orthonormal basis of  $L^2$ , write  $u = \sum_k q_k \mathbf{e}_k$  and  $\dot{u} = \sum_k \dot{q}_k \mathbf{e}_k$ , then one has the following proposition.

**Proposition 8** *The Lagrange Eqs. (8) are equivalent to*

$$\frac{\partial L}{\partial q_k} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_k} = 0 \quad (13)$$

*Proof* Taking the scalar product of (8) with  $\mathbf{e}_k$  one gets

$$\langle \mathbf{e}_k; \nabla_u L \rangle_{L^2} - \frac{d}{dt} \langle \mathbf{e}_k; \nabla_{\dot{u}} L \rangle_{L^2} = 0$$

but one has  $\langle \mathbf{e}_k; \nabla_u L \rangle = \frac{\partial L}{\partial q_k}$  and similarly for the other term. Thus the thesis follows.  $\square$

This proposition shows that, once a basis has been introduced, the Lagrange equations have the same form as in the finite dimensional case.

In the Hamiltonian case exactly the same result holds. Precisely, denoting  $v := \sum_k p_k \mathbf{e}_k$  one has the following proposition.

**Proposition 9** *The Hamilton equations of a Hamiltonian function  $H$  are equivalent to*

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}, \quad \dot{q}_k = \frac{\partial H}{\partial p_k}. \quad (14)$$

In the case of the nonlinear wave equation, in order to get a convenient form of the equations, one can choose the Fourier basis. Such a basis is defined by

$$\hat{e}_k := \begin{cases} \frac{1}{\sqrt{\pi}} \cos kx & k > 0 \\ \frac{1}{\sqrt{2\pi}} & k = 0 \\ \frac{1}{\sqrt{\pi}} \sin -kx & k < 0 \end{cases} \quad (15)$$

Thus the Hamiltonian (12) takes the form

$$H(p, q) = \sum_{k \in \mathbb{Z}} \frac{p_k^2 + \omega_k^2 q_k^2}{2} + \int_{-\pi}^{\pi} F \left( \sum_k q_k \hat{e}_k(x) \right) dx, \quad (16)$$

where  $\omega_k^2 := k^2 + \mu^2$ . For the forthcoming developments it is worth to rescale the variables by defining

$$p'_k := \frac{p_k}{\sqrt{\omega_k}}, \quad q'_k := \sqrt{\omega_k} q_k, \quad (17)$$

so that, omitting primes, the Hamiltonian takes the form

$$H(p, q) = \sum_k \omega_k \frac{p_k^2 + q_k^2}{2} + H_P(p, q) \quad (18)$$

where  $H_P$  has a zero of order higher than 2. In the following we will always study systems of the form (18). Moreover, by relabeling the variables and the frequencies it is possible to reduce the problem to the case where  $k$  varies in  $\mathbb{N} \equiv \{1, 2, 3, \dots\}$ . This is what we will assume in developing the abstract theory.

*Example 10* An example of a different nature in which the Hamiltonian takes the form (18) is the nonlinear Schrödinger equation

$$-i\dot{\psi} = \psi_{xx} + f(|\psi|^2)\psi, \quad (19)$$

where  $f$  is a smooth function. Eq. (19) has the conserved energy functional

$$H(\psi, \bar{\psi}) := \int_{-\pi}^{\pi} (|\psi|^2 + F(|\psi|^2)) dx, \quad (20)$$

where  $F$  is such that  $F' = f$ . Introduce canonical coordinates  $(p_k, q_k)$  by

$$\psi = \sum_{k \in \mathbb{Z}} \frac{p_k + iq_k}{\sqrt{2}} \hat{e}_k, \quad (21)$$

then the energy takes the form (18) with  $\omega_k = k^2$  and the NLS is equivalent to the corresponding Hamilton equations.

*Example 11* Consider the Kortweg–de Vries equation

$$u_t + u_{xxx} + uu_x = 0, \quad (22)$$

in the space of functions with zero mean value. The conserved energy is given by

$$H(u) = \int_{-\pi}^{\pi} \left( \frac{u_x^2}{2} + \frac{u^3}{6} \right) dx, \quad (23)$$

which again is also the Hamiltonian of the system. Canonical coordinates are here introduced by

$$u(x) = \sum_{k>0} \sqrt{k} (p_k \hat{e}_k + q_k \hat{e}_{-k}), \quad (24)$$

in which the Hamiltonian takes the form (18) with  $\omega_k = k^3$ .

**Remark 12** It is also interesting to study some of these equations with Dirichlet boundary conditions (DBC), typically on  $[0, \pi]$ . This will always be done by identifying the space of the functions fulfilling DBC with the space of the function fulfilling periodic boundary conditions on  $[-\pi, \pi]$  which are skew symmetric. Similarly, Neumann boundary conditions will be treated by identifying the corresponding functions with periodic even functions. In some cases (e. g. in Eq. (1) with DBC and an  $f$  which does not have particular symmetries) the equations do not extend naturally to the space of skew symmetric and this has some interesting consequences (see [7,13]).

### Basic Elements of Hamiltonian Formalism for PDEs

A suitable topology in the phase space is given by a Sobolev like topology.

For any  $s \in \mathbb{R}$ , define the Hilbert space  $\ell_s^2$  of the sequences  $x \equiv \{x_k\}_{k \geq 1}$  with  $x_k \in \mathbb{R}$  such that

$$\|x\|_s^2 := \sum_k |k|^{2s} |x_k|^2 < \infty \quad (25)$$

and the phase spaces  $\mathcal{P}_s := \ell_s^2 \oplus \ell_s^2 \equiv z \ni (p, q) \equiv (\{p_k\}, \{q_k\})$ . In  $\mathcal{P}_s$  we will sometimes use the scalar product

$$\langle (p, q), (p^1, q^1) \rangle_s := \langle p, p^1 \rangle_{\ell_s^2} + \langle q, q^1 \rangle_{\ell_s^2}. \quad (26)$$

In the following we will always assume that

$$|\omega_k| \leq C|k|^d \quad (27)$$

for some  $d$ .

**Remark 13** Defining the operator  $A_0: D(A_0) \rightarrow \mathcal{P}_s$  by  $A_0(p, q) = (\omega_k p_k, \omega_k q_k)$  one can write  $H_0 = \frac{1}{2} \langle A_0 z; z \rangle_0$ ,  $D(A_0) \supset \mathcal{P}_{s+d}$ .

Given a smooth Hamiltonian function  $\chi: \mathcal{P}_s \supset \mathcal{U}_s \rightarrow \mathbb{R}$ ,  $\mathcal{U}_s$  being an open neighborhood of the origin, we define the corresponding Hamiltonian vector field  $X_\chi: \mathcal{U}_s \mapsto \mathcal{P}_{-s}$  by

$$X_\chi \equiv \left( -\frac{\partial \chi}{\partial q_k}, \frac{\partial \chi}{\partial p_k} \right). \quad (28)$$

**Remark 14** Corresponding to a function  $\chi$  as above we will denote by  $\nabla \chi$  its gradient with respect to the  $\ell^2 \equiv \ell_0^2$  metric. Defining the operator  $J$  by  $J(p, q) := (-q, p)$  one has  $X_\chi = J \nabla \chi$ .

**Definition 15** The Poisson Bracket of two smooth functions  $\chi_1, \chi_2$  is formally defined by

$$\{\chi_1; \chi_2\} := d\chi_1 X_{\chi_2} \equiv \langle \nabla \chi_1; J \nabla \chi_2 \rangle_0. \quad (29)$$

**Remark 16** As the example  $\chi_1 = \sum_k k q_k, \chi_2 := \sum_k k p_k$  shows, there are cases where the Poisson Bracket of two functions is not defined.

For this reason a crucial role is played by the functions whose vector field is smooth.

**Definition 17** A function  $\chi \in C^\infty(\mathcal{U}_s, \mathcal{P}_s)$ ,  $\mathcal{U}_s \subset \mathcal{P}_s$  open, is said to be of class  $\text{Gen}_s$ , if the corresponding Hamiltonian vector field  $X_\chi$  is a smooth map from  $\mathcal{U}_s \rightarrow \mathcal{P}_s$ . In this case we will write  $\chi \in \text{Gen}_s$ .

**Proposition 18** Let  $\chi_1 \in \text{Gen}_s$ . If  $\chi_2 \in C^\infty(\mathcal{U}_s, \mathbb{R})$  then  $\{\chi_1, \chi_2\} \in C^\infty(\mathcal{U}_s, \mathbb{R})$ . If  $\chi_2 \in \text{Gen}_s$  then  $\{\chi_1, \chi_2\} \in \text{Gen}_s$ .

**Definition 19** A smooth coordinate transformation  $\mathcal{T}: \mathcal{P}_s \supset \mathcal{U}_s \rightarrow \mathcal{P}_s$  is said to be canonical if for any Hamiltonian function  $H$  one has  $X_{H \circ \mathcal{T}} = \mathcal{T}^* X_H \equiv d\mathcal{T}^{-1} X_H \circ \mathcal{T}$ , i. e. it transforms the Hamilton equations of  $H$  into the Hamilton equations of  $H \circ \mathcal{T}$ .

**Proposition 20** Let  $\chi_1 \in \text{Gen}_s$ , and let  $\Phi_{\chi_1}^t$  be the corresponding time  $t$  flow (which exists by standard theory). Then  $\Phi_{\chi_1}^t$  is a canonical transformation.

### Normal Form for Finite Dimensional Hamiltonian Systems

Consider a system of the form (18), but with finitely many degrees of freedom, namely a system with a Hamiltonian of the form (18) with

$$H_0(p, q) = \sum_{k=1}^n \omega_k \frac{p_k^2 + q_k^2}{2}, \quad \omega_k \in \mathbb{R} \quad (30)$$

and  $H_P$  which is a smooth function having a zero of order at least 3 at the origin.

**Definition 21** A polynomial  $Z$  will be said to be in normal form if  $\{H_0; Z\} \equiv 0$ .

**Theorem 22 (Birkhoff)** For any positive integer  $r \geq 0$ , there exist a neighborhood  $\mathcal{U}^{(r)}$  of the origin and a canonical transformation  $\mathcal{T}_r: \mathbb{R}^{2n} \supset \mathcal{U}^{(r)} \rightarrow \mathbb{R}^{2n}$  which puts the

system (18) in Birkhoff Normal Form up to order  $r$ , namely s.t.

$$H^{(r)} := H \circ \mathcal{T}_r = H_0 + Z^{(r)} + \mathcal{R}^{(r)} \quad (31)$$

where  $Z^{(r)}$  is a polynomial of degree  $r + 2$  which is in normal form,  $\mathcal{R}^{(r)}$  is small, i. e.

$$|\mathcal{R}^{(r)}(z)| \leq C_r \|z\|^{r+3}, \quad \forall z \in \mathcal{U}^{(r)}; \quad (32)$$

moreover, one has

$$\|z - \mathcal{T}_r(z)\| \leq C'_r \|z\|^2, \quad \forall z \in \mathcal{U}^{(r)}. \quad (33)$$

An inequality identical to (33) is fulfilled by the inverse transformation  $\mathcal{T}_r^{-1}$ .

If the frequencies are nonresonant at order  $r + 2$ , namely if

$$\omega \cdot K \neq 0, \quad \forall K \in \mathbb{Z}^n, \quad 0 < |K| \leq r + 2 \quad (34)$$

the function  $Z^{(r)}$  depends on the actions

$$I_j := \frac{p_j^2 + q_j^2}{2}$$

only.

**Remark 23** If the nonlinearity is analytic and the frequencies are Diophantine, i. e. there exist  $\gamma > 0$  and  $\tau$  such that

$$|\omega \cdot K| \geq \frac{\gamma}{|K|^\tau}, \quad \forall K \in \mathbb{Z}^n - \{0\}, \quad (35)$$

then one can compute the dependence of the constant  $C_r$  (cf. Eq. (32)) on  $r$  and optimize the value of  $r$  as a function of  $\|z\|$ . This allows one to improve (32) and to show that there exists and  $r_{\text{opt}}$  such that (see e. g. [40])

$$|\mathcal{R}^{(r_{\text{opt}})}(z)| \leq C \exp\left(-\frac{c}{\|z\|^{1/(\tau+1)}}\right). \quad (36)$$

In turn, such an estimate is the starting point for the proof of the celebrated Nekhoroshev's theorem [53].

The idea of the proof is to construct a canonical transformation putting the system in a form which is as simple as possible, namely the normal form. More precisely one constructs a canonical transformation  $\mathcal{T}^{(1)}$  pushing the non normalized part of the Hamiltonian to order four followed by a transformation  $\mathcal{T}^{(2)}$  pushing it to order five and so on, thus getting  $\mathcal{T}_r = \mathcal{T}^{(1)} \circ \mathcal{T}^{(2)} \circ \dots \circ \mathcal{T}^{(r)}$ . Each of the transformations  $\mathcal{T}^{(j)}$  is constructed as the time one flow of a suitable auxiliary Hamiltonian function say

$\chi_j$  (Lie transform method). It turns out that  $\chi_j$  is determined as the solution of the Homological equation

$$Z_j := \{\chi_j, H_0\} + H^{(j)} \quad (37)$$

where  $H^{(j)}$  is constructed recursively and  $Z_j$  has to be determined together with  $\chi_j$  in such a way that  $\{Z_j, H_0\} = 0$  and (37) holds. In particular  $H^{(1)}$  coincides with the first non vanishing term in the Taylor expansion of  $H_P$ .

The algorithm of solution of the Homological Eq. (37) involves division by the so called small denominators  $i\omega \cdot K$ , where  $K \in \mathbb{Z}^n - \{0\}$ , fulfills  $|K| \leq j + 2$  and  $\omega \cdot K \neq 0$ .

The above construction is more or less explicit: provided one has at disposal enough time, he can explicitly compute  $Z^{(r)}$  up to any given order. In the case of nonresonant frequencies this is not needed if one wants to understand the dynamics over long times. Indeed its features are an easy consequence of the fact that  $Z^{(r)}$  depends on the actions only. A precise statement will be given in the case of PDEs. It has to be noticed that the normal form can be used also as a starting point for the construction of invariant tori through KAM theory. To this end however one has to verify a nondegeneracy condition and this requires the explicit computation of the normal form.

In the resonant case the situation is more complicated, however, it is often enough to compute the first non vanishing term of  $Z^{(r)}$  in order to get relevant information on the dynamics. This usually requires only the ability to compute the function  $Z_1$ , defined by (37) with  $H^{(1)}$  coinciding with the first non vanishing term of the Taylor expansion of  $H_P$ . For a detailed analysis we refer to other sections of the Encyclopedia.

A particular case where one can use a coordinate independent formula for the computation of  $Z_j$  and  $\chi_j$  is the one in which the frequencies are completely resonant.

Assume that there exists  $\nu > 0$  and integer numbers  $\ell_1, \dots, \ell_n$  such that

$$\omega_k = \nu \ell_k \quad \forall k = 1, \dots, n. \quad (38)$$

**Remark 24** Denote by  $\Psi^t$  the flow of the linear system with Hamiltonian  $H_0$ , then one has

$$\Psi^{t+T} = \Psi^t, \quad T := \frac{2\pi}{\nu}, \quad t \in \mathbb{R}. \quad (39)$$

Moreover in this case one has  $\omega \cdot K \neq 0 \implies |\omega \cdot K| \geq \nu > 0$ , so there are no small denominators.

In this case one has an interesting coordinate independent formula for the solution of the homological Eq. (37).

**Lemma 25** *Let  $f$  be smooth function, defined in neighborhood of the origin. Define*

$$\begin{aligned} Z(z) &\equiv \langle f \rangle(z) := \frac{1}{T} \int_0^T f(\Psi^t(z)) dt, \\ \chi(z) &:= \frac{1}{T} \int_0^T t [f(\Psi^t(z)) - Z(\Psi^t(z))] dt, \end{aligned} \quad (40)$$

*then such quantities fulfill the equation  $\{H_0, \chi\} + f = Z$ .*

### Normal Form for Hamiltonian PDEs: General Comments

As anticipated in the introduction there are two problems in order to generalize Birkhoff's theorem to PDEs: (1) the existence of nonsmooth vector fields and (2) the appearance of small denominators accumulating at zero already at order 3.

There are two reasons why (1) is a problem. The first one is that if the vector field of  $\chi_1$  were not smooth then it would be nontrivial to ensure that it generates a flow, and thus that the normalizing transformation exists. The second related problem is that, if a transformation could be generated, then the Taylor expansion of the transformed Hamiltonian would contain a term of the form  $\{H_1; \chi_1\} = dH_1 X_{\chi_1}$ , which is typically not smooth if  $X_\chi$  is not smooth. Thus one has to show that the construction involves only functions which are of class  $\text{Gen}_s$  for some  $s$  (see Definition 17).

The difficulty related to small denominators is the following: In the finite dimensional case,  $\{\omega \cdot K \neq 0, |K| \leq r + 2\}$  implies  $|\omega \cdot K| \geq \gamma > 0$ . Thus division by  $\omega \cdot K$  is a harmless operation in the finite dimensional case. In the infinite dimensional case this is no longer true. For example, when  $\omega_k = \sqrt{k^2 + \mu^2}$  one already has

$$\inf_{0 \neq |K| \leq 3} |\omega \cdot K| = 0.$$

In order to solve such a problem one has to take advantage of a property of the nonlinearity which typically holds in PDEs and is called having *localized coefficients*. By also assuming a suitable nonresonance property for the frequency vector, one can deduce a normal form theorem identical to Theorem 22. The main difficulty consists in verifying the assumptions of the theorem. We will show how to verify such assumptions by applying this method to some typical examples.

### Normal Form for Resonant Hamiltonian PDEs and its Consequences

In the case of resonant frequencies and smooth vector field it is possible to obtain a normal form up to an exponentially small remainder.

Consider the system (18) in the phase space  $\mathcal{P}_s$  with some fixed  $s$ . Assume that the frequencies are completely resonant, namely that (38) holds (with  $k \in \mathbb{N}$ ); assume that  $H_P \in \text{Gen}_s$  and that its vector field extends to a complex analytic function in a neighborhood of the origin. Finally assume that  $H_P$  has a zero of order  $n \geq 3$  at the origin. Then we have the following theorem.

**Theorem 26 ([10,14])** *There exists a neighborhood of the origin  $\mathcal{U}_s \subset \mathcal{P}_s$  and an analytic canonical transformation  $\mathcal{T} : \mathcal{U}_s \rightarrow \mathcal{P}_s$  with the following properties:  $\mathcal{T}$  is close to identity. Namely, it satisfies*

$$\|z - \mathcal{T}(z)\|_s \leq C \|z\|_s^{n-1}. \quad (41)$$

*$\mathcal{T}$  puts the Hamiltonian in resonant normal form up to an exponentially small remainder, namely one has*

$$H \circ \mathcal{T} = H_0 + \langle H_P \rangle + Z_2 + \mathcal{R} \quad (42)$$

*where  $\langle H_P \rangle$  is the average (defined by (40)) of  $H_P$  with respect to the unperturbed flow;  $Z_2$  is in normal form, namely  $\{Z_2, H_0\} \equiv 0$ , and has a zero of order  $2n - 2$  at the origin;  $\mathcal{R}$  is an exponentially small remainder whose vector field is estimated by*

$$\|X_{\mathcal{R}}(z)\|_s \leq C \|z\|_s^{n-1} \exp\left(-\frac{C}{\|z\|_s^{n-2}}\right).$$

**Example 27** The nonlinear Schrödinger equation (19). Here one has  $\ell_k = k^2$  and  $\nu = 1$ . The Sobolev embedding theorems ensure that the vector field of the nonlinearity is analytic if  $f$  is analytic in a neighborhood of the origin. Thus Theorem 26 applies to the NLS. To deduce dynamical consequences it is convenient to explicitly compute  $\langle H_P \rangle$ . Assuming  $f(0) = 0$  and  $f'(0) = 1$  this was done in [2] using formula (40) which gives

$$\langle H_P \rangle(z) = \frac{1}{2} \left( \sum_k I_k \right)^2 - \frac{1}{8} \sum_k |I_k|^2 \quad (43)$$

where  $I_k = (p_k^2 + q_k^2)/2$  are the linear actions. Thus one has that  $H_0 + \langle H_P \rangle$  is a function of the actions only, and thus it is an integrable system. It is thus natural to study the system (42) as a perturbation of such an integrable system. This was done in [2] and [25] obtaining the results we are going to state. For simplicity we will concentrate here on

the case of Dirichlet boundary conditions, thus the function  $\psi$  will always be assumed to be skew symmetric with respect to the origin. Define

$$\epsilon_s := \left( \frac{1}{2} \int_{-\pi}^{\pi} |\partial_x^s \psi^{(0)}(x)|^2 dx \right)^{1/2}, \quad (44)$$

i. e., the  $H^s$  norm of the initial datum  $\psi^{(0)}$ ,  $s \geq 0$ , and denote by  $I_k(0)$  the initial value of the linear actions.

**Theorem 28 ([2])** *Fix  $N \geq 1$ , then there exists a constant  $\epsilon_*$ , with the property that, if the initial datum  $\psi^{(0)}$  is such that*

$$\epsilon_1 < \epsilon_*, \quad \sum_{k \geq N+1} I_k(0)^2 \leq C \epsilon_1^4 \epsilon_1^{2-1/N}, \quad (45)$$

*then along the corresponding solution of (19) one has*

$$\sum_{k \geq 1} |I_k(t) - I_k(0)|^2 \leq C' \epsilon_1^{4+1/N} \quad (46)$$

*for the times  $t$  fulfilling*

$$|t| \leq C'' \exp \left( \frac{\epsilon_*}{\epsilon_1} \right)^{1/N}.$$

This result in particular allows one to control the distance in energy norm of the solution from the torus given by the intersection of the level surfaces of the actions taken at the initial time.

**Theorem 29 ([25])** *Fix an arbitrarily large  $r$ , then there exists  $s_r$  such that for any  $s \geq s_r$  there exists  $\epsilon_{*s}$ , such that the following holds true: most of the initial data with  $\epsilon_s < \epsilon_{*s}$  give rise to solutions with*

$$\|\psi(t)\|_s \leq C \epsilon_s, \quad \forall |t| \leq \frac{C}{\epsilon_s^r}. \quad (47)$$

For the precise meaning of “most of the initial data,” we refer to the original paper. The result is based on the proof that the considered solutions remain close in the  $H^s$  topology to an infinite dimensional torus. In particular the uniform estimate of the Sobolev norm is relevant for applications to numerical analysis [30,44].

**Example 30** Consider the nonlinear wave Eq. (1) with  $\mu = 0$ . Here the frequencies are given by  $|k|$  and thus they are completely resonant. Again the smoothness of the nonlinearity is ensured by Sobolev embedding theorem. In the case of DBC in order to ensure smoothness one has also to assume that the nonlinearity is odd, namely that  $f(-u) = -f(u)$ , then Theorem 26 applies. However

in this case the computation of  $\langle H_P \rangle$  is nontrivial. It has been done (see [55]) in the case of  $f(u) = \pm u^3 + O(u^4)$  and Dirichlet boundary conditions. The result however is that the function  $\langle H_P \rangle$  does not have a particularly simple structure, and thus it is not easy to extract informations on the dynamics.

In order to extract information on the dynamics, consider the simplified system in which the remainder is neglected, namely the system with Hamiltonian

$$H_S := H_0 + \langle H_P \rangle + Z_2. \quad (48)$$

Such a system has two integrals of motion, namely  $H_0$  and  $\langle H_P \rangle + Z_2$ . Let  $\gamma_\epsilon$  be the set of the  $z$ 's at which  $\langle H_P \rangle + Z_2$  is restricted to the surface  $S_\epsilon := \{z: H_0(z) = \epsilon^2\}$  has an extremum, say a maximum. Then  $\gamma_\epsilon$  is an invariant set for the dynamics of  $H_S$ . By the invariance under the flow of  $H_0$ , one has that  $\gamma_\epsilon$  is the union of one dimensional closed curves, but generically it is just a single closed curve. In such a situation it is also a stable periodic orbit of (48) (see [35]). Actually it is very difficult to compute  $(\langle H_P \rangle + Z_2)|_{S_\epsilon}$ , but a maximum of such a function can be easily constructed by applying the implicit function theorem to a non degenerate maximum of  $\langle H_P \rangle|_{S_\epsilon}$ . The addition of the remainder then modifies the dynamics only after an exponentially long time. We are now going to state the corresponding theorem.

First remark that a critical point of  $\langle H_P \rangle|_{S_1}$  is a solution  $z_a$  of the system

$$\lambda_a A_0 z_a + \nabla \langle H_P \rangle(z_a) = 0, \quad H_0(z_a) = 1 \quad (49)$$

where we used the notations of Remarks 13 and 14. Here  $\lambda_a$  is clearly the Lagrange multiplier. The closed curve  $\gamma_a := \bigcup_t \Psi^t(z_a)$  is (the trajectory of) a periodic solution of  $H_0 + \langle H_P \rangle$ . Consider now the linear operator  $B_a := d(\nabla \langle H_P \rangle)(z_a)$ .

**Definition 31** The critical point  $z_a$  is said to be non degenerate if the system

$$\lambda_a A_0 h + B_a h = 0, \quad \langle A_0 z_a; h \rangle_0 = 0 \quad (50)$$

has at most one solution.

Under the assumptions of Theorem 32 below it is easy to prove that  $\gamma_a$  is a smooth curve and that its tangent vector  $h_a := \frac{d}{dt} \Psi^t(z_a)|_{t=0}$  is a solution of (50).

**Theorem 32 ([14,15])** *Assume that  $H_P \in \text{Gen}_s$  for any  $s$  large enough, assume also that there exists a non degenerate maximum  $z_a$  of  $\langle H_P \rangle|_{S_1}$ , then there exists a constant  $\epsilon_*$ , such that the following holds true: consider a solution  $z(t)$*



of the Hamilton equation of (18) with initial datum  $z_0$ ; if there exists  $\epsilon < \epsilon_*$ , such that

$$d_E(\epsilon \gamma_a, z_0) \leq C \epsilon^n, \quad (51)$$

then one has

$$d_E(\epsilon \gamma_a, z(t)) \leq C' \epsilon^n, \quad (52)$$

for all times  $t$  with  $|t| \leq \frac{C}{\epsilon^{r-1}} \exp\left(\frac{\epsilon_*}{\epsilon}\right)^{n-1}$ . Here  $d_E$  is the distance in the energy norm.

Such a theorem does not ensure that there exist periodic orbits of the complete system, but just a family of closed curves with the property that starting close to it one remains close to it for exponentially long times. Some results concerning the existence of true periodic orbits close to such periodic like trajectories can also be proved (see e.g. [16,19,20,42,51]).

**Example 33** In the paper [55] it has been proved that the non degeneracy assumption (50) of Theorem 32 holds for the Eq. (1) with  $f(u) = \pm u^3 + \text{higher order terms}$  and Dirichlet boundary conditions. In the case of such an equation an extremum of  $\langle H_P \rangle|_{S_1}$  is given by

$$u(x) = V_m \text{sn}(wx, m), \quad v(x) \equiv 0,$$

with  $V_m, w$  and  $m$  suitable constants, and  $\text{sn}$  the Jacobi elliptic sine. Therefore the curve  $\gamma_a$  is the phase space trajectory of the solution of the linear wave equation with such an initial datum. There are no other extrema of  $\langle H_P \rangle|_{S_1}$ . Thus the Theorem 32 ensures that solutions starting close to a rescaling of such a curve remain close to it for very long times.

## Normal Form for Nonresonant Hamiltonian PDEs

### A Statement

We turn now to the nonresonant case. The theory we will present has been developed in [6,9,43], and is closely related to the one of [4,11,37]. First we introduce the class of equations to which the theory applies. To this end it is useful to treat the  $p$ 's and the  $q$ 's exactly on an equal footing so we will denote by  $z \equiv (z_k)_{k \in \mathbb{Z}}$ ,  $\mathbb{Z} := \mathbb{Z} - \{0\}$  the set of all the variables, where

$$z_{-k} := p_k, \quad z_k := q_k \quad k \geq 1.$$

Given a polynomial function  $f: \mathcal{P}_\infty \rightarrow \mathbb{R}$  of degree  $r$  one can decompose it as follows

$$f(z) = \sum_{k_1, \dots, k_r} f_{k_1, \dots, k_r} z_{k_1} \dots z_{k_r}. \quad (53)$$

We will assume suitable localization properties for the coefficients  $f_{k_1, \dots, k_r}$  as functions of the indexes  $k_1, \dots, k_r$ .

**Definition 34** Given a multi-index  $k \equiv (k_1, \dots, k_r)$ , let  $(k_{i_1}, k_{i_2}, k_{i_3}, \dots, k_{i_r})$  be a reordering of  $k$  such that

$$|k_{i_1}| \geq |k_{i_2}| \geq |k_{i_3}| \geq \dots \geq |k_{i_r}|.$$

We define  $\mu(k) := |k_{i_1}|$  and

$$S(k) := \mu(k) + ||k_{i_1}| - |k_{i_2}||. \quad (54)$$

**Definition 35** Let  $f: \mathcal{P}_\infty \rightarrow \mathbb{R}$  be a polynomial of degree  $r$ . We say that  $f$  has *localized coefficients* if there exists  $\nu \in [0, +\infty)$  such that  $\forall N \geq 1$  there exists  $C_N$  such that for any choice of the indexes  $k_1, \dots, k_r$ , the following inequality holds:

$$|f_{k_1, \dots, k_r}| \leq C_N \frac{\mu(k)^{\nu+N}}{S(k)^N}. \quad (55)$$

**Definition 36** A function  $f \in \text{Gen}_s$  for any  $s$  large enough, is said to have *localized coefficients* if all the terms of its Taylor expansion have localized coefficients.

Some important properties of functions with localized coefficients are given by Theorem 37.

**Theorem 37** Let  $f: \mathcal{P}_\infty \rightarrow \mathbb{R}$  be a polynomial of degree  $r$  with localized coefficients, then there exists  $s_0$  such that for any  $s \geq s_0$  the vector field  $X_f$  extends to a smooth map from  $\mathcal{P}_s$  to itself; moreover the following estimate holds:

$$\|X_f(z)\| \leq C \|z\|_s \|z\|_{s_0}^{r-2}. \quad (56)$$

In particular it follows that a function with localized coefficients is of class  $\text{Gen}_s$  for any  $s \geq s_0$ .

**Theorem 38** The Poisson Bracket of two functions with localized coefficients has localized coefficients.

In order to develop perturbation theory we also need a quantitative nonresonance condition.

**Definition 39** Fix a positive integer  $r$ . The frequency vector  $\omega$  is said to fulfill the *property (r-NR)* if there exist  $\gamma > 0$ , and  $\alpha \in \mathbb{R}$  such that for any  $N$  large enough one has

$$\left| \sum_{k \geq 1} \omega_k K_k \right| \geq \frac{\gamma}{N^\alpha}, \quad (57)$$

for any  $K \in \mathbb{Z}^\infty$ , fulfilling  $0 \neq |K| := \sum_k |K_k| \leq r + 2$ ,  $\sum_{k > N} |K_k| \leq 2$ .

It is easy to see that under this condition one can solve the homological equation and that, if the known term of the equation has localized coefficients, then the solution also has localized coefficients.

**Theorem 40 ([9,12])** Fix  $r \geq 1$ , assume that the frequencies fulfill the nonresonance condition ( $r$ -NR); assume that  $H_P$  has localized coefficients. Then there exists a finite  $s_r$ , a neighborhood  $\mathcal{U}_{s_r}^{(r)}$  of the origin in  $\mathcal{P}_{s_r}$ , and a canonical transformation  $\mathcal{T} : \mathcal{U}_{s_r}^{(r)} \rightarrow \mathcal{P}_{s_r}$  which puts the system in normal form up to order  $r + 3$ , namely

$$H^{(r)} := H \circ \mathcal{T} = H_0 + Z^{(r)} + \mathcal{R}^{(r)} \quad (58)$$

where  $Z^{(r)}$  has localized coefficients and is a function of the actions  $I_k$  only;  $\mathcal{R}^{(r)}$  has a small vector field, i. e.

$$\|X_{\mathcal{R}^{(r)}}(z)\|_{s_r} \leq C \|z\|_{s_r}^{r+2}, \quad \forall z \in \mathcal{U}_{s_r}^{(r)}; \quad (59)$$

thus, one has

$$\|z - \mathcal{T}_r(z)\|_{s_r} \leq C \|z\|_{s_r}^2, \quad \forall z \in \mathcal{U}_{s_r}^{(r)}. \quad (60)$$

An inequality identical to (60) is fulfilled by the inverse transformation  $\mathcal{T}_r^{-1}$ . Finally for any  $s \geq s_r$  there exists a subset  $\mathcal{U}_s^{(r)} \subset \mathcal{U}_{s_r}^{(r)}$  open in  $\mathcal{P}_s$  such that the restriction of the canonical transformation to  $\mathcal{U}_s^{(r)}$  is analytic also as a map from  $\mathcal{P}_s \rightarrow \mathcal{P}_s$  and the inequalities (59) and (60) hold with  $s$  in place of  $s_r$ .

This theorem allows one to give a very precise description of the dynamics.

**Proposition 41** Under the same assumptions of Theorem 37,  $\forall s \geq s_r$  there exists  $\epsilon_{*s}$  such that, if the initial datum fulfills  $\epsilon := \|z_0\|_s < \epsilon_{*s}$ , then one has

$$\|z(t)\|_s \leq 4\epsilon, \quad \sum_k k^{2s} |I_k(t) - I_k(0)| \leq C\epsilon^3 \quad (61)$$

for all the times  $t$  fulfilling  $|t| \leq \epsilon^{-r}$ . Moreover there exists a smooth torus  $\mathbb{T}_0$  such that,  $\forall M \leq r$

$$d_s(z(t), \mathbb{T}_0) \leq C\epsilon^{(M+3)/2}, \quad \text{for } |t| \leq \frac{1}{\epsilon^{r-M}} \quad (62)$$

where  $d_s(\cdot, \cdot)$  is the distance in  $\mathcal{P}_s$ .

A generalization to the resonant or partially resonant case is easily obtained and can be found in [11].

### Verification of the Property of Localization of Coefficients

The property of localization of coefficients is quite abstract. We illustrate via a few examples some ways to verify it.

**Example 42** Consider the nonlinear wave Eq. (1) with Neumann boundary conditions on  $[0, \pi]$ . We recall that the corresponding space of functions will be considered as a subset of the space of periodic functions.

Consider the Taylor expansion of the nonlinearity, i. e. write  $F(u) = \sum_{r \geq 3} c_r \int_{-\pi}^{\pi} u^r$ . Then one has to prove that the functions  $f_r(u) \equiv \int_{-\pi}^{\pi} u^r$  have localized coefficients. The coefficients  $f_{k_1, \dots, k_r}$  are given by

$$f_{k_1, \dots, k_r} = \int_{-\pi}^{\pi} \cos(k_1 x) \cos(k_2 x) \dots \cos(k_r x) dx. \quad (63)$$

One could compute and estimate such a quantity directly, but it is easier to proceed in a different way: to show that  $f_3$  has localized coefficients and then to use Theorem 38 to show that each  $f_r$  has localized coefficients for any  $r$ . This is the path we will follow. Consider

$$f_{k_1, k_2, k_3} = \int_{-\pi}^{\pi} \cos(k_1 x) \cos(k_2 x) \cos(k_3 x) dx. \quad (64)$$

Since the estimate (55) is symmetric with respect to the indexes, we can assume that they are ordered,  $k_1 \geq k_2 \geq k_3$ , so that (64) =  $\pi \delta_{k_1}^{k_2+k_3}/2$ ,  $\mu(k) = k_3$ ,  $S(k) = k_3 + k_1 - k_2$  from which one immediately sees that (55) holds with  $\nu = 0$ . As a consequence one also gets that the function  $g_3 := \int_{-\pi}^{\pi} \nu u^2$  has localized coefficients. Since  $\{g_3; f_r\} = r f_{r+1}$ , by induction Theorem 38 ensures that  $f_r$  has localized coefficients for any  $r$ .

Often it is impossible to explicitly compute the coefficients  $f_{k_1, \dots, k_r}$ , so one needs a different way to verify the property.

**Example 43** Consider the nonlinear wave equation

$$u_{tt} - u_{xx} + Vu = f(u) \quad (65)$$

with Neumann boundary conditions. Here  $V$  is a smooth, even, periodic potential. The Hamiltonian reduces to the form (18) by introducing the variables  $q_k$  by  $u(x) = \sum_k q_k \varphi_k(x)$  where  $\varphi_k(x)$  are the eigenfunctions of the Sturm Liouville operator  $-\partial_{xx} + V$ , and similarly for  $\nu$ . In such a case one has

$$f_{k_1, k_2, k_3} = \int_{-\pi}^{\pi} \varphi_{k_1}(x) \varphi_{k_2}(x) \varphi_{k_3}(x) dx. \quad (66)$$

Here the idea is to consider (66) as the matrix element  $L_{k_1, k_2}$  of the operator  $L$  of multiplication by  $\varphi_{k_3}(x)$  on the basis of the eigenfunctions of the operator  $S := -\partial_{xx} + V$ . The key idea is to proceed as follows.

Let  $L$  be a linear operator which maps  $D(S^r)$  into itself for all  $r \geq 0$ , and define the sequence of operators

$$L_N := [S, L_{N-1}], \quad L_0 := L. \quad (67)$$

**Lemma 44** *Let  $S$  be as above, then, for any  $N \geq 0$  one has*

$$|L_{k_1, k_2}| = |\langle L\varphi_{k_1}; \varphi_{k_2} \rangle| \leq \frac{1}{|\lambda_{k_1} - \lambda_{k_2}|^N} |\langle L_N \varphi_{k_1}; \varphi_{k_2} \rangle| \quad (68)$$

where  $\lambda_{k_j}$  is the eigenvalue of  $S$  corresponding to  $\varphi_{k_j}$ .

Then, in order to conclude the verification of the property of localization of the coefficients, one has just to compute  $L_N$  and to estimate the scalar product product of the r.h.s. All the computations can be found in [6].

**Example 45** A third example where the verification of the property of localization of coefficients goes almost in the same way is the nonlinear Schrödinger equation

$$i\psi_t = -\psi_{xx} + V\psi + \frac{\partial F(\psi, \bar{\psi})}{\partial \bar{\psi}} \quad (69)$$

with Dirichlet Boundary conditions on  $[0, \pi]$ . Here one has to assume that  $V$  is a smooth even potential and that  $F$  is smooth and fulfills  $F(-\psi, -\bar{\psi}) = F(\psi, \bar{\psi})$  (this is required in order to leave the space of skew symmetric functions invariant, see Remark 12). Here the variables  $(p, q)$  are introduced by

$$\psi = \sum_{k \in \mathbb{Z}} \frac{p_k + iq_k}{\sqrt{2}} \varphi_k, \quad (70)$$

where  $\psi_k$  are the eigenfunctions of  $S$  with Dirichlet boundary conditions. Here the Taylor expansion of the nonlinearity has only even terms. Thus the building block for the proof of the property of localization of coefficients is the operator  $L$  of multiplication by  $\varphi_{k_3} \varphi_{k_4}$ . Then, mutatis mutandis the proof goes as in the previous case.

### Verification of the Nonresonance Property

Finally in order to apply Theorem 40 one has to verify the nonresonance property ( $r$ -NR). As usual in dynamical systems, this is done by tuning the frequencies using parameters. In the case of the nonlinear wave Eq. (1) one can use the mass  $\mu$ .

**Theorem 46 ([4,12,36])** *There exists a zero measure set  $S \subset \mathbb{R}$  such that, if  $\mu \in \mathbb{R} - S$ , then the frequencies  $\omega_k = \sqrt{k^2 + \mu^2}$ ,  $k \geq 1$  fulfill the condition ( $r$ -NR) for any  $r$ .*

Thus the Theorem 40 applies to the Eq. (1) with almost any mass.

A similar result holds for the Eq. (65), where the role of the mass is played by the average of the potential.

The situation of the nonlinear Schrödinger is more difficult. Here one can use the Fourier coefficients of the potential as parameters.

Fix  $\sigma > 0$  and, for any positive  $R$  define the space of the potentials, by

$$\mathcal{V}_R := \left\{ V(x) = \sum_{k \geq 1} v_k \cos kx \mid v'_k := v_k R^{-1} e^{\sigma k} \in \left[ -\frac{1}{2}, \frac{1}{2} \right] \text{ for } k \geq 1 \right\} \quad (71)$$

Endow such a space with the product normalized probability measure.

**Theorem 47 ([12], see also [21])** *For any  $r$  there exists a positive  $R$  and a set  $S \subset \mathcal{V}_R$  such that property ( $r$ -NR) holds for any potential  $V \in S$  and  $|\mathcal{V}_R - S| = 0$ .*

So, provided the potential is chosen in the considered set, Theorem 40 also applies to the Eq. (69).

We point out that the proof of Theorem 46 and of Theorem 47 essentially consists of two steps. First one proves that for most values of the parameters one has

$$\left| \sum_{k \geq 1} \omega_k K_k \right| \geq \frac{\gamma}{N^\alpha}, \quad \forall K \in \mathbb{Z}^\infty, \quad (72)$$

with  $0 \neq |K| := \sum_k |K_k| \leq r + 2$ .

Then one uses the asymptotic of the frequencies, namely  $\omega_k \sim ak^d$  with  $d \geq 1$ , in order to get the result.

### Non Hamiltonian PDEs

In this section we will present some results for the non Hamiltonian case.

It is useful to complexify the phase space. Thus, in this section we will always denote the space of the complex sequences  $\{z_k\}$  whose norm (defined by (25)) is finite by  $\mathcal{P}_s$ .

In the space  $\mathcal{P}_s$  consider a system of differential equations of the form

$$\dot{z}_k = \lambda_k z_k + P_k(z), \quad k \in \mathbb{Z} - \{0\} \quad (73)$$

where  $\lambda_k$  are complex numbers and  $P(z) \equiv \{P_k(z)\}$  has a zero of order at least 2 at the origin. Moreover we will assume  $P$  to be a complex analytic map from a neighborhood of the origin of  $\mathcal{P}_s$  to  $\mathcal{P}_s$ . The quantities  $\lambda_k$  are clearly the eigenvalues of the linear operator describing the linear part of the system, and for this reason they will be called “the eigenvalues”.

**Example 48** A system of the form (18) with  $H_P$  having a vector field which is analytic. The corresponding Hamilton equations have the form (73) with  $\lambda_k = -\lambda_{-k} = i\omega_k$ ,  $k \geq 1$ .

**Example 49** Consider the following nonlinear heat equation with periodic boundary conditions on  $[-\pi, \pi]$ :

$$u_t = u_{xx} - V(x)u + f(u). \quad (74)$$

If  $f$  is analytic then it can be given the form (73) by introducing the basis of the eigenfunctions  $\psi_k$  of the Sturm Liouville operator  $S := -\partial_{xx} + V$ , i.e. denoting  $u = \sum_k z_k \varphi_k$ . In this case the eigenvalues  $\lambda_k$  are the opposite of the periodic eigenvalues of  $S$ . Thus in particular one has  $\lambda_k \in \mathbb{R}$  and  $\lambda_k \sim -k^2$ .

In this context one has to introduce a suitable concept of nonresonance:

**Definition 50** A sequence of eigenvalues is said to be resonant if there exists a sequence of integer numbers  $K_k \geq 0$  and an index  $i$  such that

$$\sum_k \lambda_k K_k - \lambda_i = 0. \quad (75)$$

In the finite dimensional case the most celebrated results concerning systems of the form (73) are the Poincaré theorem, the Poincaré–Dulac theorem and the Siegel theorem [1]. The Poincaré theorem is of the form of Birkhoff's Theorem 22, while the Poincaré–Dulac and Siegel theorems guarantee (under suitable assumptions) that there exists an analytic coordinate transformation reducing the system to its normal form or linear part (no remainder!).

At present there is not a satisfactory extension of the Poincaré–Dulac theorem to PDEs (some partial results have been given in [41]). We now state a known extension of the Siegel theorem to PDEs.

**Theorem 51** ([54,63]) Assume that the eigenvalues fulfill the Diophantine type condition

$$\left| \sum_k \lambda_k K_k - \lambda_i \right| \geq \frac{\gamma}{|K|^\tau}, \quad \forall i, K \quad \text{with } 2 \leq |K|, \quad (76)$$

where  $\gamma > 0$  and  $\tau \in \mathbb{R}$  are suitable parameters; then there exists an analytic coordinate transformation defined in a neighborhood of the origin, such that the system (73) is transformed into its linear part

$$\dot{z}_k = \lambda_k z_k. \quad (77)$$

The main remark concerning this theorem is that the condition (76) is only exceptionally satisfied. If  $\lambda \in \mathbb{C}^n$  then

condition (76) is generically satisfied only if  $\tau > (n-2)/2$ . Nevertheless some examples where such an equation is satisfied are known [54].

The formalism of Sect. “Normal Form for Hamiltonian Nonresonant PDEs” can be easily generalized to the non Hamiltonian case giving rise to a generalization of Poincaré's theorem, which we are going to state.

Given a polynomial map  $P: \mathcal{P}_\infty \rightarrow \mathcal{P}_{-\infty}$  one can expand it on the canonical basis  $\mathbf{e}_k$  of  $\mathcal{P}_0$  as follows:

$$P(z) = \sum_{k_1, \dots, k_r, i} P_{k_1, \dots, k_r, i}^i z_{k_1} \dots z_{k_r} \mathbf{e}_i, \quad P_{k_1, \dots, k_r, i}^i \in \mathbb{C}. \quad (78)$$

**Definition 52** A polynomial map  $P$  is said to have localized coefficients if there exists  $\nu \in [0, +\infty)$  such that  $\forall N \geq 1$  there exists  $C_N$  such that for any choice of the indexes  $k_1, \dots, k_r, i$  following the inequality holds:

$$\left| P_{k_1, \dots, k_r, i}^i \right| \leq C_N \frac{\mu(k, i)^{\nu+N}}{S(k, i)^N}, \quad (79)$$

where  $(k, i) = (k_1, \dots, k_r, i)$ . A map is said to have localized coefficients if for any  $s$  large enough, it is a smooth map from  $\mathcal{P}_s$  to itself and each term of its Taylor expansion has localized coefficients.

**Definition 53** The eigenvalues are said to be strongly nonresonant at order  $r$  if for any  $N$  large enough, any  $K = (K_{k_1}, \dots, K_{k_r})$  and any index  $i$  such that  $|K| \leq r$  and there are at most two of the indexes  $k_1, \dots, k_r, i$  larger than  $N$  the following inequality holds:

$$\left| \sum_k \lambda_k K_k - \lambda_i \right| \geq \frac{\gamma}{N^\alpha}. \quad (80)$$

**Theorem 54** Assume that the nonlinearity has localized coefficients and that the eigenvalues are strongly nonresonant at order  $r$ , then there exists an  $s_r$  and an analytic coordinate transformation  $\mathcal{T}_r: \mathcal{U}_{s_r} \rightarrow \mathcal{P}_{s_r}$  which transforms the system (73) to the form

$$\dot{z}_k = \lambda_k z_k + \mathcal{R}_k(z), \quad (81)$$

where the following inequality holds

$$\|\mathcal{R}(z)\|_{s_r} \leq C \|z\|_{s_r}^r. \quad (82)$$

## Extensions and Related Results

The theory presented here applies to quite general semilinear equations in one space dimension. At present a satisfactory theory applying to quasilinear equations and/or equations in more than one space dimensions is not available. The main difficulty for the extension of the theory to

semilinear equations in higher space dimension is related to the nonresonance condition. The general theory can be easily extended to the case where the differences  $\omega_k - \omega_l$  between frequencies accumulate only at a set constituted by isolated points.

*Example 55* Consider the nonlinear wave equation on the  $d$ -dimensional sphere

$$u_{tt} - \Delta_g u + \mu^2 u = f(x, u), \quad x \in S^d \quad (83)$$

with  $\Delta_g$  the Laplace Beltrami operator; the frequencies are given by  $\omega_k = \sqrt{k(k+d-1) + \mu^2}$  and their differences accumulate only at integers. A version of Theorem 40 applicable to (83) was proved in [9]. As a consequence in particular one has a lower bound on the existence time  $t$  of the small amplitude solutions of the form  $|t| \geq \epsilon^{-r}$ , where  $\epsilon$  is proportional to a high Sobolev norm of the initial datum. An extension to  $u_{tt} - \Delta_g u + Vu = f(x, u)$  is also known.

*Example 56* A similar result was proved in [11] for the nonlinear Schrödinger equation

$$-i\dot{\psi} = -\Delta\psi + V * \psi + f(|\psi|^2)\psi, \quad x \in \mathbb{T}^d, \quad (84)$$

where the star denotes convolution.

The only general result available at present for quasilinear system is the following theorem.

**Theorem 57 ([5])** Fix  $r \geq 1$ , assume that the frequency vector fulfills condition (72) and that there exists  $d_1$  such that the vector field of  $H_P$  is a smooth map from  $\mathcal{P}_{s+d_1}$  to  $\mathcal{P}_s$  for any  $s$  large enough. Then the same result of Theorem 40 holds, but the functions do not necessarily have localized coefficients and, for any  $s$  large enough **the remainder is estimated by**

$$\|X_{\mathcal{R}^{(r)}}(z)\|_s \leq C\|z\|_{s+d_r}^{r+2}, \quad \forall z \in \mathcal{U}_s^{(r)}; \quad (85)$$

where  $d_r$  is a large positive number.

The estimate (85) shows that the remainder is small only when considered as an operator extracting a lot of derivatives. In particular it is non trivial to use such a theorem in order to extract information on the dynamics. Following the approach of [8] and [5] this can be done using the normal form to construct approximate solutions and suitable versions of the Gronwall lemma to compare it with solutions of the true system. This however allows to control the dynamics only over times of the order of  $\epsilon^{-1}$ ,  $\epsilon$  being again a measure of the size of the initial datum. Such a theory has been applied to quasilinear wave equations in [5] and to the Fermi Pasta Ulam problem in [17].

Among the large number of papers containing related results we recall [41,47,56,59]. A stronger result for the nonlinear wave equation valid over times of order  $\epsilon^{-2}$  can be found in [36].

## Future Directions

Future directions of research include both purely theoretical aspects and applications to other sciences.

From a purely theoretical point of view, the most important open problems pertain to the validity of normal form theory for equations in which the nonlinearity involves derivatives, and for general equations in more than one space dimension.

These results would be particularly important since the kind of equations appearing in most domains of physics are quasilinear and higher dimensional.

Concerning applications, we would like to describe a few of them which would be of interest.

- **Water wave problem.** The problem of description of the free surface of a fluid has been shown to fit in the scheme of Hamiltonian dynamics [62]. Normal form theory could allow one to extract the relevant informations on the dynamics in different situations [31,38], ranging from the theory of Tsunamis [32] to the theory of fluid interface, which is relevant e. g. to the construction of oil platforms [33].
- **Quantum mechanics.** A Bose condensate is known to be well described by the Gross Pitaevskii equation. When the potential is confining, such an equation is of the form (18). Normal form theory has already been used for some preliminary results [18], but a systematic investigation could lead to interesting new results.
- **Electromagnetic theory and magnetohydrodynamics.** The equations have a Hamiltonian form; normal form theory could help to describe some instability arising in plasmas.
- **Elastodynamics.** Here one of the main theoretical open problems is the stability of equilibria which are a minimum of the energy. The problem is that in higher dimensions the conservation of energy does not ensure enough smoothness of the solution to ensure stability. Such a problem is of the same kind as the one solved in [9] when dealing with the existence times of the nonlinear wave equation.

## Bibliography

1. Arnold V (1984) Chapitres supplémentaires de la théorie des équations différentielles ordinaires. Mir, Moscow
2. Bambusi D (1999) Nekhoroshev theorem for small amplitude solutions in nonlinear Schrödinger equation. Math Z 130:345–387



3. Bambusi D (1999) On the Darboux theorem for weak symplectic manifolds. *Proc Amer Math Soc* 127(11):3383–3391
4. Bambusi D (2003) Birkhoff normal form for some nonlinear PDEs. *Comm Math Phys* 234:253–283
5. Bambusi D (2005) Galerkin averaging method and Poincaré normal form for some quasilinear PDEs. *Ann Sc Norm Super Pisa Cl Sci* 4(5):669–702
6. Bambusi D (2008) A Birkhoff normal form theorem for some semilinear PDEs. In: Craig W (ed) *Hamiltonian dynamical systems and applications*. Springer
7. Bambusi D, Carati A, Penati T (2007) On the relevance of boundary conditions for the FPU paradox. Preprint *Institt Lombardo Accad Sci Lett Rend A* (to appear)
8. Bambusi D, Carati A, Ponno A (2002) The nonlinear Schrödinger equation as a resonant normal form. *DCDS-B* 2:109–128
9. Bambusi D, Delort JM, Grébert B, Szeftel J (2007) Almost global existence for Hamiltonian semi-linear Klein–Gordon equations with small Cauchy data on Zoll manifolds. *Comm Pure Appl Math* 60(11):1665–1690
10. Bambusi D, Giorgilli A (1993) Exponential stability of states close to resonance in infinite-dimensional Hamiltonian systems. *J Statist Phys* 71(3–4):569–606
11. Bambusi D, Grebert B (2003) Forme normale pour NLS en dimension quelconque. *Compt Rendu Acad Sci Paris* 337:409–414
12. Bambusi D, Grébert B (2006) Birkhoff normal form for partial differential equations with tame modulus. *Duke Math J* 135(3):507–567
13. Bambusi D, Muraro D, Penati T (2008) Numerical studies on boundary effects on the FPU paradox. *Phys Lett A* 372(12):2039–2042
14. Bambusi D, Nekhoroshev NN (1998) A property of exponential stability in the nonlinear wave equation close to main linear mode. *Phys D* 122:73–104
15. Bambusi D, Nekhoroshev NN (2002) Long time stability in perturbations of completely resonant PDE's, Symmetry and perturbation theory. *Acta Appl Math* 70(1–3):1–22
16. Bambusi D, Paleari S (2001) Families of periodic orbits for resonant PDE's. *J Nonlinear Sci* 11:69–87
17. Bambusi D, Ponno A (2006) On metastability in FPU. *Comm Math Phys* 264(2):539–561
18. Bambusi D, Sacchetti A (2007) Exponential times in the one-dimensional Gross–Pitaevskii equation with multiple well potential. *Commun Math Phys* 234(2):136
19. Berti M, Bolle P (2003) Periodic solutions of nonlinear wave equations with general nonlinearities. *Commun Math Phys* 243:315–328
20. Berti M, Bolle P (2006) Cantor families of periodic solutions for completely resonant nonlinear wave equations. *Duke Math J* 134(2):359–419
21. Bourgain J (1996) Construction of approximative and almost-periodic solutions of perturbed linear Schrödinger and wave equations. *Geom Funct Anal* 6:201–230
22. Bourgain J (1996) On the growth in time of higher Sobolev norms of smooth solutions of Hamiltonian PDE. *Int Math Res Not* 6:277–304
23. Bourgain J (1997) On growth in time of Sobolev norms of smooth solutions of nonlinear Schrödinger equations in  $\mathbb{R}^D$ . *J Anal Math* 72:299–310
24. Bourgain J (1998) Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equation. *Ann Math* 148:363–439
25. Bourgain J (2000) On diffusion in high-dimensional Hamiltonian systems and PDE. *J Anal Math* 80:1–35
26. Bourgain J (2005) Green's function estimates for lattice Schrödinger operators and applications. In: *Annals of Mathematics Studies*, vol 158. Princeton University Press, Princeton
27. Bourgain J (2005) On invariant tori of full dimension for 1D periodic NLS. *J Funct Anal* 229(1):62–94
28. Bourgain J, Kaloshin V (2005) On diffusion in high-dimensional Hamiltonian systems. *J Funct Anal* 229(1):1–61
29. Chernoff PR, Marsden JE (1974) Properties of infinite dimensional Hamiltonian systems In: *Lecture Notes in Mathematics*, vol 425. Springer, Berlin
30. Cohen D, Hairer E, Lubich C (2008) Long-time analysis of nonlinearly perturbed wave equations via modulated Fourier expansions. *Arch Ration Mech Anal* 187(2):341–368
31. Craig W (1996) Birkhoff normal forms for water waves In: *Mathematical problems in the theory of water waves* (Luminy, 1995), *Contemp Math*, vol 200. Amer Math Soc, Providence, RI, pp 57–74
32. Craig W (2006) Surface water waves and tsunamis. *J Dyn Differ Equ* 18(3):525–549
33. Craig W, Guyenne P, Kalisch H (2005) Hamiltonian long-wave expansions for free surfaces and interfaces. *Comm Pure Appl Math* 58(12):1587–1641
34. Craig W, Wayne CE (1993) Newton's method and periodic solutions of nonlinear wave equations. *Comm Pure Appl Math* 46:1409–1498
35. Dell'Antonio GF (1989) Fine tuning of resonances and periodic solutions of Hamiltonian systems near equilibrium. *Comm Math Phys* 120(4):529–546
36. Delort JM, Szeftel J (2004) Long-time existence for small data nonlinear Klein–Gordon equations on tori and spheres. *Int Math Res Not* 37:1897–1966
37. Delort J-M, Szeftel J (2006) Long-time existence for semi-linear Klein–Gordon equations with small Cauchy data on Zoll manifolds. *Amer J Math* 128(5):1187–1218
38. Dyachenko AI, Zakharov VE (1994) Is free-surface hydrodynamics an integrable system? *Phys Lett A* 190:144–148
39. Eliasson HL, Kuksin SB (2006) KAM for non-linear Schroedinger equation. *Ann of Math*. Preprint (to appear)
40. Fassò F (1990) Lie series method for vector fields and Hamiltonian perturbation theory. *Z Angew Math Phys* 41(6):843–864
41. Foias C, Saut JC (1987) Linearization and normal form of the Navier–Stokes equations with potential forces. *Ann Inst H Poincaré Anal Non Linéaire* 4:1–47
42. Gentile G, Mastropietro V, Procesi M (2005) Periodic solutions for completely resonant nonlinear wave equations with Dirichlet boundary conditions. *Comm Math Phys* 256(2):437–490
43. Grébert B (2007) Birkhoff normal form and Hamiltonian PDEs. *Partial differential equations and applications*, 1–46 *Sémin Congr*, 15 Soc Math France, Paris
44. Hairer E, Lubich C (2006) Conservation of energy, momentum and actions in numerical discretizations of nonlinear wave equations
45. Kappeler T, Pöschel J (2003) *KdV & KAM*. Springer, Berlin
46. Klainerman S (1983) On almost global solutions to quasilinear

wave equations in three space dimensions. *Comm Pure Appl Math* 36:325–344

47. Krol MS (1989) On Galerkin–averaging method for weakly nonlinear wave equations. *Math Meth Appl Sci* 11:649–664
48. Kuksin SB (1987) Hamiltonian perturbations of infinite-dimensional linear systems with an imaginary spectrum. *Funct Anal Appl* 21:192–205
49. Kuksin SB (1993) Nearly integrable infinite-dimensional Hamiltonian systems. Springer, Berlin
50. Kuksin SB, Pöschel J (1996) Invariant Cantor manifolds of quasi-periodic oscillations for a nonlinear Schrödinger equation. *Ann Math* 143:149–179
51. Lidskij BV, Shulman EI (1988) Periodic solutions of the equation  $u_{tt} - u_{xx} + u^3 = 0$ . *Funct Anal Appl* 22:332–333
52. Marsden J (1972) Darboux's theorem fails for weak symplectic forms. *Proc Amer Math Soc* 32:590–592
53. Nekhoroshev NN (1977) Exponential estimate of the stability of near integrable Hamiltonian systems. *Russ Math Surv* 32(6): 1–65
54. Nikolenko NV (1986) The method of Poincaré normal form in problems of integrability of equations of evolution type. *Russ Math Surv* 41:63–114
55. Paleari S, Bambusi D, Cacciatori S (2001) Normal form and exponential stability for some nonlinear string equations. *ZAMP* 52:1033–1052
56. Pals H (1996) The Galerkin–averaging method for the Klein–Gordon equation in two space dimensions. *Nonlinear Anal TMA* 27:841–856
57. Pöschel J (2002) On the construction of almost-periodic solutions for a nonlinear Schrödinger equation. *Ergod Th Dyn Syst* 22:1–22
58. Soffer A, Weinstein MI (1999) Resonances, radiation damping and instability in Hamiltonian nonlinear wave equations. *Invent Math* 136(1):9–74
59. Stroucken ACJ, Verhulst F (1987) The Galerkin–averaging method for nonlinear, undamped continuous systems. *Math Meth Appl Sci* 335:520–549
60. Wayne CE (1990) Periodic and quasi-periodic solutions of nonlinear wave equations via KAM theory. *Comm Math Phys* 127:479–528
61. Weinstein A (1969) Symplectic structures on Banach manifolds. *Bull Amer Math Soc* 75:1040–1041
62. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. *Appl Mech Tech Phys* 2:190–194
63. Zehnder E (1978) C L Siegel's linearization theorem in infinite dimensions. *Manuscripta Math* 23:363–371

## Perturbation Theory in Quantum Mechanics

LUIGI E. PICASSO<sup>1,2</sup>, LUCIANO BRACCI<sup>1,2</sup>,  
EMILIO D'EMILIO<sup>1,2</sup>

<sup>1</sup> Dipartimento di Fisica, Università di Pisa, Pisa, Italy

<sup>2</sup> Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, Pisa, Italy

## Article Outline

Glossary

Definition of the Subject

Introduction

Presentation of the Problem and an Example

Perturbation of Point Spectra: Nondegenerate Case

Perturbation of Point Spectra: Degenerate Case

The Brillouin–Wigner Method

Symmetry and Degeneracy

Problems with the Perturbation Series

Perturbation of the Continuous Spectrum

Time Dependent Perturbations

Future Directions

Bibliography

## Glossary

**Hilbert space** A Hilbert space  $\mathcal{H}$  is a normed complex vector space with a Hermitian scalar product. If  $\varphi, \psi \in \mathcal{H}$  the scalar product between  $\varphi$  and  $\psi$  is written as  $(\varphi, \psi) \equiv (\psi, \varphi)^*$  and is taken to be linear in  $\psi$  and antilinear in  $\varphi$ : if  $a, b \in \mathbb{C}$ , the scalar product between  $a\varphi$  and  $b\psi$  is  $a^*b(\varphi, \psi)$ . The norm of  $\psi$  is defined as  $\|\psi\| \equiv \sqrt{(\psi, \psi)}$ . With respect to the norm  $\|\cdot\|$ ,  $\mathcal{H}$  is a complete metric space. In the following  $\mathcal{H}$  will be assumed to be separable, that is any complete orthonormal set of vectors is countable.

**States and observables** In quantum mechanics the states of a system are represented as vectors in a Hilbert space  $\mathcal{H}$ , with the convention that proportional vectors represent the same state. Physicists mostly use Dirac's notation: the elements of  $\mathcal{H}$  are represented by  $|\cdot\rangle$  ("ket") and the scalar product between  $|\varphi\rangle$  and  $|\psi\rangle$  is written as  $\langle\varphi|\psi\rangle$  ("bracket"). The observables, i. e. the physical quantities that can be measured, are represented by linear Hermitian (more precisely: self-adjoint) operators on  $\mathcal{H}$ . The eigenvalues of an observable are the only possible results of the measurement of the observable. The observables of a system are generally the same of the corresponding classical system: energy, angular momentum, etc., i. e. they are of the form  $f(q, p)$ , with  $q \equiv (q_1, \dots, q_n)$ ,  $p \equiv (p_1, \dots, p_n)$  the position and momentum canonical variables of the system:  $q_i$  and  $p_i$  are observables, i. e. operators, which satisfy the commutation relations  $[q_i, q_j] \equiv q_i q_j - q_j q_i = 0$ ,  $[p_i, p_j] = 0$ ,  $[q_i, p_j] = i\hbar \delta_{ij}$ , with  $\hbar$  the Planck's constant  $h$  divided by  $2\pi$ .

**Representations** Since separable Hilbert spaces are isomorphic, it is always possible to represent the elements of  $\mathcal{H}$  as elements of  $l_2$ , the space of the sequences  $\{u_i\}$ ,  $u_i \in \mathbb{C}$ , with the scalar product

$(v, u) \equiv \sum_i v_i^* u_i$ . This can be done by choosing an orthonormal basis of vectors  $e_i$  in  $\mathcal{H} : (e_i, e_j) = \delta_{ij}$  and defining  $u_i = (e_i, u)$ ; with Dirac's notations  $|A\rangle \rightarrow \{a_i\}$ ,  $a_i = \langle e_i | A \rangle$ . Linear operators  $\xi$  are then represented by  $\{\xi_{ij}\}$ ,  $\xi_{ij} = (e_i, \xi e_j) \equiv \langle e_i | \xi | e_j \rangle$ . The  $\xi_{ij}$  are called "matrix elements" of  $\xi$  in the representation  $e_i$ . If  $\xi^\dagger$  is the Hermitian-conjugate of  $\xi$ , then  $(\xi^\dagger)_{ij} = \xi_{ji}^*$ . If the  $e_i$  are eigenvectors of  $\xi$  then the (infinite) matrix  $\xi_{ij}$  is diagonal, the diagonal elements being the eigenvalues of  $\xi$ .

**Schrödinger representation** A different possibility is to represent the elements of  $\mathcal{H}$  as elements of  $L^2[\mathbb{R}^n]$ , the space of the square-integrable functions on  $\mathbb{R}^n$ , where  $n$  is the number of degrees of freedom of the system. This can be done by assigning how the operators  $q_i$  and  $p_i$  act on the functions of  $L^2[\mathbb{R}^n]$ : in the Schrödinger representation the  $q_i$  are taken to act as multiplication by  $x_i$  and the  $p_i$  as  $-i\hbar\partial/\partial x_i$ : if  $|A\rangle \rightarrow \psi_A(x_1, \dots, x_n)$ , then

$$\begin{aligned} q_i |A\rangle &\rightarrow x_i \psi_A(x_1, \dots, x_n), \\ p_i |A\rangle &\rightarrow -i\hbar \partial \psi_A(x_1, \dots, x_n) / \partial x_i. \end{aligned}$$

**Schrödinger equation** Among the observables, the Hamiltonian  $H$  plays a special role. It determines the time evolution of the system through the time dependent Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = H\psi,$$

and its eigenvalues are the energy levels of the system. The eigenvalue equation  $H\psi = E\psi$  is called the Schrödinger equation.

### Definition of the Subject

In the investigation of natural phenomena a crucial role is played by the comparison between theoretical predictions and experimental data. Those practicing the two arts of the trade continuously put challenges to one another either presenting data which ask for an explanation or proposing new experimental verifications of a theory. Celestial mechanics offers the first historical instance of this interplay: the elliptical planetary orbits discovered by Kepler were explained by Newton; when discrepancies from the elliptical paths definitely emerged it was necessary to add the effects of the heavier planets to the dominant role of the sun, until persistent discrepancies between theory and experiment asked for the drastic revision of the theory of gravitation put forth by Einstein, a revision which in turn offered a lot of new effects to observe, some of which have been verified only recently.

In this dialectic interaction between theory and experiment only the simplest problem, that of a planet moving in the field of the sun within Newton's theory, can be solved exactly. All the rest was calculated by means of perturbation theory. Generally speaking, perturbation theory is the technique of finding an approximate solution to a problem where to a dominant factor, which allows for an exact solution (zeroth order solution), other "perturbing" factors are added which are outweighed by the dominant factor and are expected to bring small corrections to the zeroth order solution.

Perturbation theory is ever-pervasive in physics, but an area where it plays a major role is quantum mechanics. In the early days of this discipline, the interpretation of atomic spectra was made possible only by a heavy use of perturbation theory, since the only exactly soluble problem was that of the hydrogen atom without external fields. The explanation of the Stark spectra (hydrogen in a constant electric field) and of the Zeeman spectra (atom in a magnetic field) was only possible when a perturbation theory tailored to the Schrödinger equation, which rules the atomic world, was devised. As for heavier atoms, in no case an exact solution for the Schrödinger equation is available: they could only be treated as a perturbation of simpler "hydrogenoid" atoms. Most of the essential aspects of atomic and molecular physics could be explained quantitatively in a few years by recourse to suitable forms of perturbation theory. Not only did it explain the position of the spectral lines, but also their relative intensities, and the absence of some lines which showed the impossibility of the corresponding transitions (selection rules) found a convincing explanation when symmetry considerations were introduced. When later more accurate measurements revealed details in the hydrogen spectrum (the Lamb shift) that only quantum field theory was able to explain, perturbation theory gained a new impetus which sometimes resulted in the anticipation of theory (quantum electrodynamics) over experiment as to the accuracy of the effect to be measured.

An attempt to describe all the forms that perturbation theory assumes in the various fields of physics would be vain. We will limit to illustrate its role and its methods in quantum mechanics, which is perhaps the field where it has reached its most mature development and finds its widest applications.

### Introduction

An early example of the use of perturbation theory which clearly illustrates its main ideas is offered by the study of

the free fall of a body [29]. The equation of motion is

$$\dot{\vec{v}} = \vec{g} + 2\vec{v} \times \vec{\Omega} + \vec{\Omega} \times (\vec{r} \times \vec{\Omega}) \quad (1)$$

where  $\vec{g}$  is the constant gravity acceleration and  $\vec{\Omega}$  the angular velocity of the rotation of the earth about its axis.  $\Omega$  is the parameter characterizing the perturbation. If we wish to find the eastward deviation of the trajectory to first order in  $\Omega$  we can neglect the third term in the RHS of Eq. (1), whose main effect is to cause a southward deviation (in the northern hemisphere). The ratio of the second term to the first one in the RHS of Eq. (1) (the effective perturbation parameter) is  $\Omega \sqrt{h/g} \simeq 10^{-4}$  for the fall from a height  $h \sim 100$  m, so we can find the effect of  $\Omega$  by writing  $\vec{v} = \vec{v}_0 + \vec{v}_1$  in Eq. (1), where  $\vec{v}_0$  is the zeroth order solution ( $\vec{v}_0 = \vec{g}t$  if  $\vec{v}_0(0) = 0$ ) and  $\vec{v}_1$  obeys

$$\dot{\vec{v}}_1 = 2\vec{v}_0 \times \vec{\Omega} = 2t\vec{g} \times \vec{\Omega}. \quad (2)$$

The solution is  $\vec{r} = \vec{h} + \frac{1}{2}\vec{g}t^2 + \frac{1}{3}t^3\vec{g} \times \vec{\Omega}$ . The eastward deviation is the deviation in the direction of  $\vec{g} \times \vec{\Omega}$  and its value is  $\delta = \frac{1}{3}\vec{t}^3 g \Omega \cos \theta$ , where  $\theta$  is the latitude and  $\vec{t}$  the zeroth order time of fall,  $\vec{t} = \sqrt{2gh}$ .

While the above example is a nice illustration of the main features of perturbation theory (identification of a perturbation parameter whose powers classify the contributions to the solution, existence of a zeroth order exact solution) the beginning of modern perturbation theory can be traced back to the work of Rayleigh on the theory of sound [33]. In essence, he wondered how the normal modes of a vibrating string

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial^2 v}{\partial x^2}, \quad v(0, t) = v(\pi, t) = 0 \quad (3)$$

are modified when passing from a constant density  $\rho = 1$  to a perturbed density  $\rho + \epsilon \sigma(x)$ . To solve this problem he wrote down most of the formulae [10,33] which are still in use to calculate the first order correction to non-degenerate energy levels in quantum mechanics.

The equation for the normal modes is

$$u''(x) + \lambda \rho(x) u(x) = 0, \quad u(0) = u(\pi) = 0. \quad (4)$$

Let  $u_n^{(0)} \equiv \sqrt{2/\pi} \sin nx$  be the unperturbed solution for the  $n$ th mode,  $\lambda_n = n^2$ , and  $u_n^{(0)} + \epsilon u_n^{(1)}$  the perturbed solution through first order, corresponding to a frequency  $\lambda_n + \epsilon \mu_n$ . By writing the equation for  $u_n^{(1)}$

$$\frac{d^2 u_n^{(1)}}{dx^2} + \lambda_n u_n^{(1)} + \mu_n u_n^{(0)} + \lambda_n \sigma u_n^{(0)} = 0, \quad u_n^{(1)}(0) = u_n^{(1)}(\pi) = 0 \quad (5)$$

after multiplying by  $u_r^{(0)}$  and using Green's theorem he found

$$\mu_n = -\lambda_n \int_0^\pi \sigma(x) (u_n^{(0)})^2 dx, \quad (6)$$

$$a_{rn} \equiv \int_0^\pi u_r^{(0)} u_n^{(1)} dx = \frac{\lambda_n}{\lambda_r - \lambda_n} \int_0^\pi \sigma u_r^{(0)} u_n^{(0)} dx \quad (r \neq n) \quad (7)$$

$$\int_0^\pi u_n^{(0)} u_n^{(1)} dx = 0. \quad (8)$$

As an application Rayleigh found the position  $\pi/2 + \delta x \equiv \pi/2 + \epsilon \tau$  of the nodal point of the perturbed mode  $n = 2$  when the perturbation to the density is  $\sigma = \kappa \delta(x - \pi/4)$ . The vanishing of  $u_2^{(0)} + \epsilon u_2^{(1)}$  determines  $2\sqrt{2/\pi} \tau = u_2^{(1)}(\pi/2)$ . By Eq. (7) the function  $u_2^{(1)}$  has an expansion  $u_2^{(1)} = \sum_{n \neq 2} a_{n2} u_n^{(0)}$ ,  $a_{n2} = \frac{4\kappa}{n^2 - 4} \sin n\pi/4$ . The result for  $\tau$  is

$$\tau = -\frac{2\kappa}{\pi\sqrt{2}} \left( 1 + \frac{1}{3} - \frac{1}{5} - \frac{1}{7} + \frac{1}{9} + \frac{1}{11} - \dots \right) = -\frac{\kappa}{2}.$$

(The series in brackets is equal to  $\int_0^1 (1+x^2)/(1+x^4) dx = 1/2 \int_0^\infty (1+x^2)/(1+x^4) dx$ , which can be calculated by contour integration.)

Perturbation theory was revived by Schrödinger, who introduced it into quantum mechanics in a pioneering work of 1926 [44]. There, he applied the concepts and methods which Rayleigh had put forth to the case where the zeroth order problem was a partial differential equation with non-constant coefficients, and he wrote down, in the language of wave mechanics, all the relevant formulae which yield the correction to the energy levels and to the wave functions for the case of both non-degenerate and degenerate energy levels. As an application he calculated the shift of the energy levels of the hydrogen atom in a constant electric field by two different methods. First he observed that in parabolic coordinates the wave equation is separable also with a constant electric field, which implies that in the subspace of the states with equal zeroth order energy the perturbation is diagonal in the basis of the parabolic eigenfunctions, thus circumventing the intricacies of the degenerate case. Later, he used the spherical coordinates, which entails a non diagonal perturbation matrix and calls for the full machinery of the perturbation theory for degenerate eigenvalues.

It is of no use to repeat here Schrödinger's calculations, since the methods which they use are at the core of modern perturbation theory, which is referred to as the

Rayleigh–Schrödinger (RS) perturbation theory. It rapidly superseded other approaches (as that by Born, Heisenberg and Jordan [7], who worked in the framework of the matrix quantum mechanics), and will be presented in the following sections.

### Presentation of the Problem and an Example

The most frequent application of perturbation theory in quantum mechanics is the approximate calculation of point spectra. The Hamiltonian  $H$  is split into an exactly solvable part  $H_0$  (the unperturbed Hamiltonian) plus a term  $V$  (the perturbation) which, in a sense to be specified later, is small with respect to  $H_0$ :  $H = H_0 + V$ . In many cases the perturbation contains an adjustable parameter which depends on the actual physical setting. For example, for a system in an external field this parameter is the field strength. For weak fields one expects the spectrum of  $H$  to differ only slightly from the spectrum of  $H_0$ . In these cases it is convenient to single out the dependence on a parameter by setting

$$H(\lambda) \equiv H_0 + \lambda V. \quad (9)$$

Accordingly we will write the Schrödinger equation as

$$H(\lambda)\psi(\lambda) = E(\lambda)\psi(\lambda). \quad (10)$$

We will retain the form Eq. (9) of the Hamiltonian even when  $H$  does not contain a variable parameter, thereby understanding that the actual eigenvalues and eigenvectors are the values at  $\lambda = 1$ .

The basic idea of the RS perturbation theory is that the eigenvalues and eigenvectors of  $H$  can be represented as power series

$$\psi(\lambda) = \sum_0^\infty \lambda^n \psi^{(n)}, \quad E(\lambda) = \sum_0^\infty \lambda^n \epsilon_n, \quad (11)$$

whose coefficients are determined by substituting expansions Eq. (11) into Eq. (10) and equating terms of equal order in  $\lambda$ . Generally, only the first few terms of the series can be explicitly computed, and the primary task of the RS perturbation theory is their calculation. The practicing scientist who uses perturbation theory never has to tackle the mathematical problem of the convergence of the series. This problem, however, or more generally the connection between the truncated perturbation sums and the actual values of the energy and the wave function, is fundamental for the consistency of perturbation theory and will be touched upon in a later section.

Before expounding the technique of the RS perturbation theory we will consider a simple (two-dimensional)

problem which can be solved exactly, since in its discussion several features of perturbation theory will emerge clearly, concerning both the behavior of the energy  $E(\lambda)$  and the behavior of the Taylor expansion of this function. From the physical point of view a system with two-dimensional Hilbert space  $\mathbb{C}^2$  can be thought of as a particle with spin 1/2 when the translational degrees of freedom are ignored.

Let us write the Hamiltonian  $H = H_0 + \lambda V$  in a representation where  $H_0$  is diagonal:

$$\begin{aligned} H &= \begin{pmatrix} E_1^0 & 0 \\ 0 & E_2^0 \end{pmatrix} + \lambda \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \\ &= \begin{pmatrix} E_1^0 + \lambda V_{11} & \lambda V_{12} \\ \lambda V_{12}^* & E_2^0 + \lambda V_{22} \end{pmatrix}. \end{aligned} \quad (12)$$

We consider first the case  $E_1^0 \neq E_2^0$ ,  $V_{12} \neq 0$ . The exact eigenvalues  $E_{1,2}(\lambda)$  of  $H$  are found by solving the secular equation:

$$\begin{aligned} E_{1,2}(\lambda) &= \frac{1}{2} \left[ (E_1^0 + \lambda V_{11}) + (E_2^0 + \lambda V_{22}) \pm \sqrt{\Delta(\lambda)} \right], \\ \Delta(\lambda) &\equiv ((E_1^0 + \lambda V_{11}) - (E_2^0 + \lambda V_{22}))^2 + 4\lambda^2 |V_{12}|^2. \end{aligned} \quad (13)$$

$$\Delta(\lambda) \equiv ((E_1^0 + \lambda V_{11}) - (E_2^0 + \lambda V_{22}))^2 + 4\lambda^2 |V_{12}|^2. \quad (14)$$

The corresponding eigenvectors, in the so called intermediate normalization defined by  $(\psi(0), \psi(\lambda)) = 1$ , are

$$\psi_1(\lambda) = \left( 1, \frac{\sqrt{\Delta(\lambda)} - (E_1^0 - E_2^0) - \lambda(V_{11} - V_{22})}{2\lambda V_{12}} \right) \quad (15)$$

$$\psi_2(\lambda) = \left( -\frac{\sqrt{\Delta(\lambda)} - (E_1^0 - E_2^0) - \lambda(V_{11} - V_{22})}{2\lambda V_{21}}, 1 \right). \quad (16)$$

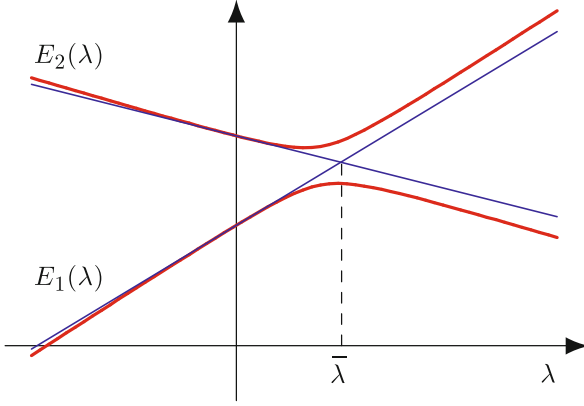
Expanding  $E_{1,2}(\lambda)$  through order  $\lambda^3$  we get:

$$\begin{aligned} E_1(\lambda) &= E_1^0 + \lambda V_{11} + \lambda^2 \frac{|V_{12}|^2}{E_1^0 - E_2^0} \\ &\quad - \lambda^3 \frac{|V_{12}|^2 (V_{11} - V_{22})}{(E_1^0 - E_2^0)^2} + O(\lambda^4) \end{aligned} \quad (17)$$

$$\begin{aligned} E_2(\lambda) &= E_2^0 + \lambda V_{22} - \lambda^2 \frac{|V_{12}|^2}{E_1^0 - E_2^0} \\ &\quad + \lambda^3 \frac{|V_{12}|^2 (V_{11} - V_{22})}{(E_1^0 - E_2^0)^2} + O(\lambda^4). \end{aligned} \quad (18)$$

At order 1 only the diagonal matrix elements of  $V$  contribute to  $E_{1,2}$ . The validity of the approximation requires  $\lambda |V_{12}| \ll |E_1^0 - E_2^0|$ . If this condition is not satisfied, that is if the eigenvalues  $E_1^0, E_2^0$  are “quasi-degenerate”, all terms of the expansion can be numerically of the





**Perturbation Theory in Quantum Mechanics, Figure 1**  
The behavior of the exact eigenvalues  $E_{1,2}(\lambda)$  when  $V_{12} = 0$  (blue lines) and when  $V_{12} \neq 0$  (red lines)

same order of magnitude and no approximation of finite order makes sense.

Note that, within the first order approximation, “level crossing” ( $E_1(\lambda) = E_2(\lambda)$ ) occurs at

$$\bar{\lambda} = -(E_1^0 - E_2^0) / (V_{11} - V_{22}). \quad (19)$$

On the other hand Eq. (13) shows that level-crossing is impossible, unless  $V_{12} = 0$ , in which case the first order approximation yields the exact result. If  $V_{12} \neq 0$  the behavior of the levels  $E_1(\lambda)$  and  $E_2(\lambda)$  near  $\bar{\lambda}$  is shown in Fig. 1: the two levels “repel” each other [49].

At first order the eigenvectors  $\psi_{1,2}(\lambda)$  are

$$\psi_1^{[1]} = (1, -\lambda V_{21} / (E_2^0 - E_1^0)) \quad (20)$$

$$\psi_2^{[1]} = (-\lambda V_{12} / (E_1^0 - E_2^0), 1). \quad (21)$$

The expectation value  $(\psi_1^{[1]}, H\psi_1^{[1]}) / (\psi_1^{[1]}, \psi_1^{[1]})$  of the Hamiltonian over  $\psi_1^{[1]}$ , for example, is

$$E_1^0 + \lambda V_{11} + \lambda^2 \frac{|V_{12}|^2}{E_1^0 - E_2^0} - \lambda^3 \frac{|V_{12}|^2 (V_{11} - V_{22})}{(E_1^0 - E_2^0)^2} - \lambda^4 \frac{|V_{12}|^4}{(E_1^0 - E_2^0)^3} + O(\lambda^5) \quad (22)$$

which agrees with  $E_1(\lambda)$  up to the  $\lambda^3$  terms (the correct fourth order term contains also  $|V_{12}|^2 (V_{11} - V_{22})^2 / (E_1^0 - E_2^0)^3$ ). This is an example of Wigner’s  $(2n + 1)$ -theorem [52], see Subsect. “Wigner’s Theorem”.

The power expansions of  $E_{1,2}(\lambda)$  and  $\psi_{1,2}(\lambda)$  converge in the disk  $|\lambda| < |E_1^0 - E_2^0| / \sqrt{(V_{11} - V_{22})^2 + 4|V_{12}|^2}$ . The denominator is just twice the infimum over  $a$  of the operator norm of  $V - aI$ . Since adding to  $V$  a multiple of

the identity does not affect the convergence properties of the Taylor’s series of  $E(\lambda)$ , we see that the convergence domain always contains the disk  $|\lambda| < |E_1^0 - E_2^0| / 2\|V\|$ , a property which holds true for any bounded perturbation in Hilbert space (see Sect. “Problems with the Perturbation Series”).

If  $H_0$  is degenerate, that is  $E_1^0 = E_2^0 \equiv E^0$ , then the eigenvalues are obtained by diagonalizing  $V$ . The degeneracy is removed and the corrections to the eigenvalues are of first order in  $\lambda$ :

$$E_{1,2}(\lambda) = E^0 + \frac{1}{2}\lambda(V_{11} + V_{22}) \pm \sqrt{(V_{11} - V_{22})^2 + 4|V_{12}|^2} \quad (23)$$

while the eigenvectors are  $\lambda$  independent.

The infinite dimensional case is much more involved. In particular, in most cases the perturbation series does not converge at all, that is its radius of convergence vanishes. However, we shall meet again the three situations discussed above: the case of non-degenerate eigenvalues  $E_n^0$  such that  $|E_n^0 - E_m^0| \gg |\lambda V_{nm}|$ , the case of degenerate eigenvalues and finally the case of “quasi-degenerate” eigenvalues, i. e. groups of eigenvalues  $E_{n_i}^0$  such that  $|E_{n_i}^0 - E_{n_j}^0| \lesssim |\lambda V_{n_i n_j}|$ . As discussed above, in this last case  $H_{n_i n_j}$  must be diagonalized exactly prior to applying perturbation theory.

### Perturbation of Point Spectra: Nondegenerate Case

In this section we consider an eigenvector  $\psi_0$  of  $H_0$  belonging to a non-degenerate eigenvalue  $E_0$  and apply the RS theory to determine the power expansions Eq. (11) such that Eq. (10) is satisfied, the Hamiltonian  $H(\lambda)$  being given by Eq. (9). The case of a degenerate eigenvalue will be considered in Sect. “Perturbation of Point Spectra: Degenerate Case”. For both cases the starting point is the substitution of the expansions Eq. (11) into Eq. (10), which, upon equating terms with equal powers, yields the following system of equations

$$(H_0 - E_0)\psi^{(n)} + V\psi^{(n-1)} = \sum_{k=0}^{n-1} \psi^{(k)} \epsilon_{n-k}, \quad n = 1, 2, \dots \quad (24)$$

A perturbative calculation of the energy and the wave function through order  $h$  amounts to calculating  $\epsilon_n$  and  $\psi^{(n)}$  up to  $n = h$  and truncating the series in Eq. (11) at  $n = h$ .

### Corrections to the Energy and the Eigenvectors

In the following let  $\psi_k$ ,  $E_k$  be the normalized eigenvectors and the eigenvalues of  $H_0$ , and let  $\Delta E_{k0} \equiv E_k - E_0$ ,  $V_{hk} \equiv (\psi_h, V\psi_k)$ . The correction  $\epsilon_n$  is recursively defined in terms of the lower order corrections to the energy and the wave function: by left multiplying Eq. (24) by  $\psi_0$  we find

$$\epsilon_n = (\psi_0, V\psi^{(n-1)}) - \sum_{h=1}^{n-1} (\psi_0, \psi^{(h)}) \epsilon_{n-h}. \quad (25)$$

Similarly, the components  $(\psi_k, \psi^{(n)})$ ,  $k \neq 0$ , are found by left-multiplying by  $\psi_k$ ,  $k \neq 0$ :

$$\begin{aligned} (\psi_k, \psi^{(n)}) &= -(\psi_k, V\psi^{(n-1)}) \Delta E_{k0}^{-1} \\ &+ \sum_{h=1}^{n-1} (\psi_k, \psi^{(h)}) \epsilon_{n-h} \Delta E_{k0}^{-1}. \end{aligned} \quad (26)$$

Note that, even if the functions  $\psi^{(k)}$ 's for  $k < n$  were known, still  $(\psi_0, \psi^{(n)})$  is intrinsically undefined, since to any solution of Eq. (24) we are allowed to add any multiple of  $\psi_0$ . The reason of this indeterminacy is that Eq. (10) defines  $\psi(\lambda)$  only up to a multiplicative factor  $\alpha(\lambda)$ . Even the normalization condition  $(\psi(\lambda), \psi(\lambda)) = 1$  leaves  $\psi(\lambda)$  undetermined by a phase factor  $\exp(i\varphi(\lambda))$ ,  $\varphi(\lambda) \in \mathbb{R}$ . On the contrary, the corrections  $\epsilon_n$  as well as all the expectation values (up to order  $n$ ) are unaffected by these modifications of the wave function  $\psi(\lambda)$  [16].

We can turn to our advantage the indeterminacy of  $(\psi_0, \psi^{(n)})$  by requiring that in the expression of  $\epsilon_n$ , Eq. (25), the dependence on the values of  $(\psi_0, \psi^{(k)})$ ,  $k \leq n-1$ , disappears. For example, after writing

$$\begin{aligned} (\psi_0, V\psi^{(n-1)}) &= V_{00} (\psi_0, \psi^{(n-1)}) + \sum_{h \neq 0} V_{0h} (\psi_h, \psi^{(n-1)}) \end{aligned}$$

the independence of  $(\psi_0, \psi^{(n-1)})$  implies  $\epsilon_1 = V_{00}$ . Next, requiring  $\epsilon_n$  to be independent of  $(\psi_0, \psi^{(n-2)})$  determines  $\epsilon_2$  and so on, until finally Eq. (25) gives  $\epsilon_n$ . As an example we carry through this procedure for  $n = 3$ . Starting from

$$\begin{aligned} \epsilon_3 &= V_{00} (\psi_0, \psi^{(2)}) + \sum_{k \neq 0} V_{0k} (\psi_k, \psi^{(2)}) \\ &- \epsilon_1 (\psi_0, \psi^{(2)}) - \epsilon_2 (\psi_0, \psi^{(1)}) \end{aligned}$$

we first find

$$\epsilon_1 = V_{00}. \quad (27)$$

Next, from Eq. (26) for  $n = 2$ , we get

$$\begin{aligned} \epsilon_3 &= - \sum_{k \neq 0} \frac{|V_{0k}|^2}{\Delta E_{k0}} (\psi_0, \psi^{(1)}) \\ &- \sum_{h,k \neq 0} \frac{V_{0k}}{\Delta E_{k0}} V_{kh} (\psi_h, \psi^{(1)}) \\ &+ \sum_{k \neq 0} \frac{V_{0k}}{\Delta E_{k0}} (\psi_k, \psi^{(1)}) \epsilon_1 - \epsilon_2 (\psi_0, \psi^{(1)}). \end{aligned}$$

The independence of  $(\psi_0, \psi^{(1)})$  implies

$$\epsilon_2 = - \sum_{k \neq 0} \frac{|V_{0k}|^2}{\Delta E_{k0}}. \quad (28)$$

Finally, by using Eq. (26) for  $n = 1$  we find

$$\epsilon_3 = \sum_{h,k \neq 0} \frac{V_{0k}}{\Delta E_{k0}} \frac{V_{kh}}{\Delta E_{h0}} V_{h0} - \epsilon_1 \sum_{k \neq 0} \frac{|V_{0k}|^2}{\Delta E_{k0}^2}. \quad (29)$$

Note that if  $\epsilon_n$  is required, the lower order corrections being known, one can use a simplified version of Eqs. (25) and (26) where the terms  $(\psi_0, \psi^{(k)})$  are omitted since the beginning. Once the values of  $\epsilon_k$ ,  $k \leq n$ , have been determined, Eq. (26) yields  $(\psi_k, \psi^{(n)})$ . For example, for the first order correction to the wave function we have

$$(\psi_k, \psi^{(1)}) = - \frac{V_{k0}}{\Delta E_{k0}}. \quad (30)$$

By suitably choosing the arbitrary factor  $\alpha(\lambda)$  we referred to after Eq. (26) we can impose  $(\psi_0, \psi(\lambda)) = 1$ . With this choice (known as the “intermediate normalization”, since  $\psi(\lambda)$  is not normalized) we have  $(\psi_0, \psi^{(k)}) = 0$  for any  $k > 0$ . As a result, for the wave function through order  $n$  we find

$$\psi^{[n]} \equiv \psi_0 + \sum_{k=1}^n \lambda^k \psi^{(k)} \equiv \psi_0 + \delta_n \psi \quad (31)$$

with

$$(\psi_0, \psi^{[n]}) = 1. \quad (32)$$

Using the intermediate normalization the expression of  $\epsilon_n$  is

$$\epsilon_n = (\psi_0, V\psi^{(n-1)}), \quad (33)$$

while the value of  $(\psi_k, \psi^{(n-1)})$  can be read immediately in the expression of  $\epsilon_n$ :  $(\psi_k, \psi^{(n-1)})$  is obtained from  $\epsilon_n$  by omitting in each term the factor  $V_{0k}$  and the sum over  $k$ .

For example the wave function  $\psi^{[2]} \equiv \psi_0 + \lambda \psi^{(1)} + \lambda^2 \psi^{(2)}$  in the intermediate normalization by Eqs. (28) and (29) is

$$\begin{aligned} \psi^{[2]} = & \psi_0 - \lambda \sum_{k=1} \psi_k \frac{V_{k0}}{\Delta E_{k0}} \\ & + \lambda^2 \sum_{h,k=1} \psi_k \frac{V_{kh}}{\Delta E_{k0}} \frac{V_{h0}}{\Delta E_{h0}} \\ & - \lambda^2 \epsilon_1 \sum_{k=1} \psi_k \frac{V_{k0}}{\Delta E_{k0}^2} . \end{aligned} \quad (34)$$

In order to calculate expectation values, transition probabilities and so on one needs the normalized wave function

$$\psi_N^{[n]} = N^{1/2} (\psi_0 + \delta_n \psi) \quad (35)$$

with  $N^{-1} = 1 + (\delta_n \psi, \delta_n \psi)$ .  $N$  can be chosen real. Note that the wave function  $\psi^{[1]}$  is correctly normalized up to first order.

From the above equations one sees in which sense the perturbation  $V$  must be small with respect to the unperturbed Hamiltonian  $H_0$ : the separation between the unperturbed energy levels must be large with respect to the matrix elements of the perturbation between those levels and the total correction  $\delta E$  to  $E_0$  should be small with respect to  $|E_i - E_0|$ ,  $E_i$  standing for any other level of the spectrum of  $H_0$ .

### Wigner's Theorem

From Eq. (24) it follows that

$$H\psi^{[n]} = E^{[n]}\psi^{[n]} + O(\lambda^{n+1}) ,$$

whence one should infer that, if  $E$  is the exact energy,  $E - (\psi_N^{[n]}, H\psi_N^{[n]}) = O(\lambda^{n+1})$ . It is therefore remarkable Wigner's result that the knowledge of  $\psi^{[n]}$  allows the calculation of the energy up to order  $2n + 1$  (Wigner's  $2n + 1$  theorem) [52]. Indeed, he proved that, if  $E$  is the exact energy,

$$E - \frac{(\psi^{[n]}, H\psi^{[n]})}{(\psi^{[n]}, \psi^{[n]})} = O(\lambda^{2n+2}) .$$

To this purpose, let

$$\chi^{(n+1)} = \psi - \frac{\psi^{[n]}}{\sqrt{(\psi^{[n]}, \psi^{[n]})}}$$

where  $\psi$  is the normalized exact wave function,  $H\psi = E\psi$ . Then

$$\begin{aligned} \chi^{(n+1)} &= O(\lambda^{n+1}) , \\ (\psi, \chi^{(n+1)}) + (\chi^{(n+1)}, \psi) \\ &= -(\chi^{(n+1)}, \chi^{(n+1)}) = O(\lambda^{2n+2}) . \end{aligned}$$

As a consequence

$$\frac{(\psi^{[n]}, H\psi^{[n]})}{(\psi^{[n]}, \psi^{[n]})} - (\psi, H\psi) = O(\lambda^{2n+2}) .$$

We make explicit this point with an example. Since

$$\psi_N^{[1]} = \frac{\psi_0 + \lambda \psi^{(1)}}{\sqrt{1 + \lambda^2 (\psi^{(1)}, \psi^{(1)})}} ,$$

by using Eq. (30) and recalling Eq. (28) and Eq. (29) we have

$$\begin{aligned} (\psi_N^{[1]}, H\psi_N^{[1]}) &= E_0 + \lambda \epsilon_1 + \frac{\lambda^2 \epsilon_2 + \lambda^3 \epsilon_3}{1 + \lambda^2 (\psi^{(1)}, \psi^{(1)})} \\ &= E_0 + \lambda \epsilon_1 + \lambda^2 \epsilon_2 + \lambda^3 \epsilon_3 + O(\lambda^4) . \end{aligned}$$

### The Feynman–Hellmann Theorem

The RS perturbative expansion rests on the hypothesis that both the eigenvalues  $E(\lambda)$  and the corresponding eigenvectors  $\psi(\lambda)$  admit a power series expansion, in short, that they are analytic functions of  $\lambda$  in a neighborhood of the origin. As we shall see in Sect. “Problems with the Perturbation Series”, as a rule it is not so and the perturbative expansion gives rise only to a formal series. For this reason it is advisable to derive the various terms of the perturbation expansion without assuming analyticity. If we need  $E(\lambda)$  and  $\psi(\lambda)$  through order  $n$  it is sufficient to assume that, as functions of  $\lambda$ , they are  $C^{n+1}$ , that is continuously differentiable  $(n + 1)$  times. The procedure consists in taking the derivatives of Eq. (10) [15,28]: at the first step we get

$$H\psi'(\lambda) + V\psi(\lambda) = E'(\lambda)\psi(\lambda) + E(\lambda)\psi'(\lambda) \quad (36)$$

and by left multiplication by  $\psi(\lambda)$ , with  $(\psi(\lambda), \psi(\lambda)) = 1$ , we get

$$E'(\lambda) = (\psi(\lambda), V\psi(\lambda)) , \quad (37)$$

which is a special case of the Feynman–Hellmann theorem [17,23]:

$$\frac{\partial E}{\partial \lambda} = \left( \psi(\lambda), \frac{\partial H}{\partial \lambda} \psi(\lambda) \right) . \quad (38)$$

For  $\lambda = 0$  we find

$$E'(0) = V_{00} , \quad (39)$$

whence  $\epsilon_1 = V_{00}$ , in agreement with Eq. (27). Next, after left multiplying Eq. (36) by  $\psi_k$  and taking  $\lambda = 0$  we get

$$(\psi_k, \psi') = -\frac{V_{k0}}{\Delta E_{k0}} \quad (40)$$

which, again, agrees with Eq. (30). Taking now the derivative of (37) at  $\lambda = 0$  and using Eq. (40) we obtain

$$E''(0) = 2 \sum_{k=1} V_{0k} (\psi_k, \psi') = -2 \sum_{k=1} \frac{|V_{0k}|^2}{\Delta E_{k0}} \quad (41)$$

whence  $\epsilon_2 = \frac{1}{2}E''(0)$ , in agreement with Eq. (28).

It is clear that the procedure can be pursued to any allowed order, and that the results for the energy corrections, as well as for the wave functions, are the same we obtained earlier by the RS technique. However, the conceptual difference, that no analyticity hypothesis is required, is important since in many cases this hypothesis is not satisfied.

As to the relation of  $E^{[n]} \equiv E_0 + \lambda \epsilon_1 + \dots + \lambda^n \epsilon_n$  with  $E(\lambda)$  we recall that, since by assumption  $E(\lambda)$  is  $C^{n+1}$ , we can write Taylor's formula with a remainder:

$$E(\lambda) = \sum_0^n \frac{E^{(p)}}{p!} \lambda^p + \frac{E^{(n+1)}(\theta\lambda)}{(n+1)!} \lambda^{n+1}, \quad 0 < \theta < 1. \quad (42)$$

As observed in [28], since for small  $\lambda$  the sign of the remainder is the sign of  $E^{(n+1)}(0)\lambda^{n+1}$ , Eq. (42) allows to establish whether the sum in Eq. (42) underestimates or overestimates  $E(\lambda)$ . Moreover, if two consecutive terms, say  $q$  and  $q+1$ , have opposite sign, then (for sufficiently small  $\lambda$ )  $E(\lambda)$  is bracketed between the partial sums including and excluding the  $q$ th term. It is a pity that no one can anticipate how small such a  $\lambda$  should be. (Of course these remarks apply to the RS truncated series as well.)

### Perturbation of Point Spectra: Degenerate Case

The case when the unperturbed energy  $E_0$  is a degenerate eigenvalue of  $H_0$ , i.e. in the Hilbert space there exists a subspace  $W_0$  generated by a set  $\{\psi_0^{(i)}\}$ ,  $1 \leq i \leq n_0$ , of orthogonal normalized states, such that each  $\psi_0$  in  $W_0$  obeys  $(H_0 - E_0)\psi_0 = 0$ , deserves a separate treatment. The main problem is that, if  $\psi(\lambda)$  is an eigenstate of the exact Hamiltonian  $H = H_0 + \lambda V$ , we do not know beforehand which state of  $W_0$   $\psi(0)$  is.

In order to use a more compact notation it is convenient to introduce the projection  $P_0$  onto the subspace  $W_0$  and its complement  $Q_0$

$$P_0 \psi = \sum_{i=0}^{n_0} \psi_0^{(i)} (\psi_0^{(i)}, \psi), \quad Q_0 \equiv I - P_0, \quad (43)$$

where  $\psi_0^{(i)}$ ,  $1 \leq i \leq n_0$ , is any orthonormal basis of  $W_0$ . The Hamiltonian  $H = H_0 + \lambda V$  can be written as

$$\begin{aligned} H &= (P_0 + Q_0)(H_0 + \lambda V)(P_0 + Q_0) \\ &= E_0 P_0 + \lambda V_{PP} + \lambda V_{PQ} + \lambda V_{QP} + \lambda V_{QQ} \\ &\quad + Q_0 H_0 Q_0, \end{aligned} \quad (44)$$

where

$$\begin{aligned} V_{PP} &= P_0 V P_0, & V_{PQ} &= P_0 V Q_0, \\ V_{QP} &= Q_0 V P_0, & V_{QQ} &= Q_0 V Q_0. \end{aligned} \quad (45)$$

After projecting the Schrödinger equation onto  $W_0$  and its orthogonal complement  $W_0^\perp$ , we find

$$(E_0 + \lambda V_{PP})P_0 \psi + \lambda V_{PQ} Q_0 \psi = E P_0 \psi \quad (46)$$

$$\lambda V_{QP} P_0 \psi + Q_0 H_0 Q_0 \psi + \lambda V_{QQ} Q_0 \psi = E Q_0 \psi. \quad (47)$$

Letting

$$H_{QQ} = Q_0 H Q_0 = Q_0 H_0 Q_0 + \lambda V_{QQ} \quad (48)$$

$Q_0 \psi$  can be extracted from Eq. (47):

$$Q_0 \psi = \lambda (E - H_{QQ})^{-1} V_{QP} P_0 \psi. \quad (49)$$

Note that in Eq. (47) the operator  $H_{QQ}$  acts on vectors of  $W_0^\perp$  and that  $E - H_{QQ}$  does possess an inverse in  $W_0^\perp$ . Indeed, the existence of a vector  $\zeta$  in  $W_0^\perp$  such that

$$(H_{QQ} - E)\zeta = 0 \quad (50)$$

contradicts the assumptions which perturbation theory is grounded in: the separation between  $E(\lambda)$  and  $E(0)$  should be negligible with respect to the separation between different eigenvalues of  $H_0$ . Actually, if  $\psi_k$  is such that  $H_0 \psi_k = E_k \psi_k$ ,  $E_k \neq E_0$ , by left multiplying Eq. (50) by  $\psi_k$  we would find

$$(E - E_k)(\psi_k, \zeta) = \lambda(\psi_k, V\zeta)$$

where the LHS is of order 0 in  $\lambda$ , whereas the RHS of order 1.

By substituting Eq. (49) into Eq. (46) we have

$$\begin{aligned} (E_0 + \lambda V_{PP})P_0 \psi + \lambda^2 V_{PQ}(E - H_{QQ})^{-1} V_{QP} P_0 \psi \\ = E P_0 \psi. \end{aligned} \quad (51)$$

The energy shifts  $\Delta E \equiv E - E_0$  appear as eigenvalues of an operator  $A(E)$  acting in  $W_0$

$$A(E) \equiv \lambda V_{PP} + \lambda^2 V_{PQ}(E - H_{QQ})^{-1} V_{QP} \quad (52)$$

which however still depends on the unknown exact energy  $E$ . A calculation of the energy corrections up to a given order is possible, starting from Eq. (52), provided we expand the term  $(E - H_{QQ})^{-1}$  as far as is necessary to include all terms of the requested order.

### Corrections to the Energy and the Eigenvectors

The contributions  $\epsilon_i$  are extracted from Eq. (51) by expanding

$$E = E_0 + \lambda \epsilon_1 + \lambda^2 \epsilon_2 + \dots$$

$$P_0 \psi = \varphi_0 + \lambda \varphi_1 + \lambda^2 \varphi_2 + \dots$$

and equating terms of equal order. At the first order, since the second term in the LHS of Eq. (51) is of order 2 or larger, we have

$$V_{PP} \varphi_0 = \epsilon_1 \varphi_0. \quad (53)$$

The first order corrections to the energy are the eigenvalues of the matrix  $V_{PP}$  and the corresponding zeroth order wave function is the corresponding eigenvector.

In the most favorable case the eigenvalues of  $V_{PP}$  are simple, and the degeneracy is completely removed since the first order of perturbation theory. In this case, in order to get the higher order corrections, we can avail ourselves of the arbitrariness in the way of splitting the exact Hamiltonian into a solvable unperturbed Hamiltonian plus a perturbation by putting

$$H = (H_0 + \lambda V_{PP}) + \lambda(V - V_{PP}) \equiv H'_0 + \lambda V'. \quad (54)$$

The eigenvectors of  $H'_0$  are the solutions of Eq. (53), with eigenvalues  $E_0 + \lambda \epsilon_1^{(i)}$ ,  $1 \leq i \leq n_0$ , plus the eigenvectors  $\psi_j$  of  $H_0$  with eigenvalues  $E_j \neq E_0$ . Since the eigenvalues  $E_0 + \lambda \epsilon_1^{(i)}$  are no longer degenerate, the formalism of non-degenerate perturbation theory can be applied, but a warning is in order. When in higher perturbation orders a denominator  $\Delta E_{k0}$  occurs with the index  $k$  referring to another vector of the basis of  $W_0$ , this denominator is of order  $\lambda$  and consequently the order of the term containing this denominator is lower than the naive  $V$ -counting would imply. In each such term, the effective order is the  $V$ -counting order minus the number of these denominators. As shown below, this situation occurs starting from terms of order 4 in the perturbation  $V$ . Note that, also in the case of non-complete removal of the degeneracy, the procedure outlined above, with obvious modifications, can be applied to search the higher order corrections to those eigenvalues which at first order turn out to be non-degenerate.

If a residual degeneracy still exists, i. e. an eigenvalue  $\epsilon_1$  of Eq. (53) is not simple, we must explore the higher order corrections until the degeneracy, if possible, is removed. First of all we must disentangle the contributions of different order in  $\lambda$  from  $(E - H_{QQ})^{-1}$ . Since

$$E - H_{QQ} = (E - Q_0 H_0 Q_0) \cdot [1 - \lambda(E - Q_0 H_0 Q_0)^{-1} V_{QQ}] ,$$

we have

$$(E - H_{QQ})^{-1} = \sum_0^\infty \lambda^n [(E - Q_0 H_0 Q_0)^{-1} V_{QQ}]^n \cdot (E - Q_0 H_0 Q_0)^{-1}. \quad (55)$$

As the energy  $E$  still contains contributions of any order, the operator  $(E - Q_0 H_0 Q_0)^{-1}$  must in turn be expanded into a series in  $\lambda$ . To make notations more readable, we define

$$\frac{Q_0}{a^n} \equiv (E_0 - Q_0 H_0 Q_0)^{-n}. \quad (56)$$

The second order terms from Eqs. (51) and (55) give

$$V_{PP} \varphi_1 + V_{PQ} \frac{Q_0}{a} V_{QP} \varphi_0 = \epsilon_2 \varphi_0 + \epsilon_1 \varphi_1. \quad (57)$$

Let  $P_0^{(i)}$  be the projections onto the subspaces  $W_0^{(i)}$  of  $W_0$  corresponding to the eigenvalues  $\epsilon_1^{(i)}$ :

$$P_0 = \sum_i P_0^{(i)}, \quad V_{PP} = \lambda \sum_i \epsilon_1^{(i)} P_0^{(i)}, \quad (58)$$

$$P_1 \equiv P_0^{(1)}, \quad \epsilon_1 \equiv \epsilon_1^{(1)}.$$

By projecting onto  $W_1 \equiv W_0^{(1)}$  and recalling that  $\varphi_0$  is in  $W_1$  we get

$$P_1 V_{PQ} \frac{Q_0}{a} V_{QP} \varphi_0 = \epsilon_2 \varphi_0, \quad (59)$$

whence  $\epsilon_2$  is an eigenvalue of the operator

$$V_1 \equiv P_1 V_{PQ} \frac{Q_0}{a} V_{QP} P_1 = P_1 V \frac{Q_0}{a} V P_1. \quad (60)$$

Again, if the eigenvalue  $\epsilon_2$  is non-degenerate, we can use the previous theory by splitting the Hamiltonian as

$$H = (H_0 + \lambda V_{PP} + \lambda V_1) + \lambda(V - V_{PP} - V_1) \equiv H''_0 + \lambda V''. \quad (61)$$

The vectors which make  $V_1$  diagonal belong to non-degenerate eigenvalues of  $H''_0$ , hence the non-degenerate theory can be applied. If, on the contrary, the eigenvalue  $\epsilon_2$  of  $V_1$  is still degenerate, the above procedure can be carried out one step further, with the aim of removing the residual degeneracy. We work out the calculation for  $\epsilon_3$ , since a new aspect of degenerate perturbation theory emerges: a truly third order term which is the ratio of a term of order 4 in the potential and a term of first order (see Eq. (67) below).



From Eq. (51) and (55) we extract the contribution of order 3:

$$\begin{aligned} V_{PP}\varphi_2 + V_{PQ} \frac{Q_0}{a} V_{QP}\varphi_1 - \epsilon_1 V_{PQ} \frac{Q_0}{a^2} V_{QP}\varphi_0 \\ + V_{PQ} \frac{Q_0}{a} V_{QQ} \frac{Q_0}{a} V_{QP}\varphi_0 = \epsilon_1 \varphi_2 + \epsilon_2 \varphi_1 + \epsilon_3 \varphi_0. \end{aligned} \quad (62)$$

We want to convert this equation into an eigenvalue problem for  $\epsilon_3$ . In analogy with Eq. (58) we have

$$\begin{aligned} P_1 = \sum_i P_1^{(i)}, \quad V_1 = \sum_i \epsilon_2^{(i)} P_1^{(i)}, \\ P_2 \equiv P_1^{(1)}, \quad \epsilon_2 \equiv \epsilon_2^{(1)}. \end{aligned} \quad (63)$$

Since  $P_2 V_{PP} = \epsilon_1 P_2$ , first we eliminate  $\varphi_2$  by applying  $P_2$  to Eq. (62):

$$\begin{aligned} P_2 V_{PQ} \frac{Q_0}{a} V_{QP}\varphi_1 - \epsilon_1 P_2 V_{PQ} \frac{Q_0}{a^2} V_{QP}\varphi_0 \\ + P_2 V_{PQ} \frac{Q_0}{a} V_{QQ} \frac{Q_0}{a} V_{QP}\varphi_0 = \epsilon_3 P_2 \varphi_0 + \epsilon_2 P_2 \varphi_1. \end{aligned} \quad (64)$$

Writing  $\varphi_1 = \sum_i P_0^{(i)} \varphi_1$ , since  $P_2 V_1 = \epsilon_2 P_2$  the contribution with  $i = 1$  of the first term in the LHS of Eq. (64) is  $P_2 V_1 \varphi_1 = \epsilon_2 P_2 \varphi_1$ . Hence, Eq. (64) reads

$$\begin{aligned} \sum_{i \neq 1} P_2 V_{PQ} \frac{Q_0}{a} V_{QP} P_0^{(i)} \varphi_1 - \epsilon_1 P_2 V_{PQ} \frac{Q_0}{a^2} V_{QP}\varphi_0 \\ + P_2 V_{PQ} \frac{Q_0}{a} V_{QQ} \frac{Q_0}{a} V_{QP}\varphi_0 = \epsilon_3 P_2 \varphi_0 + \epsilon_2 \sum_{i \neq 1} P_0^{(i)} \varphi_1. \end{aligned} \quad (65)$$

Finally,  $P_0^{(i)} \varphi_1$ ,  $i \neq 1$ , is extracted from Eq. (57) by projecting with  $P_0^{(i)}$ ,  $i \neq 1$ , and recalling that  $P_0^{(i)} \varphi_0 = 0$  if  $i \neq 1$ :

$$P_0^{(i)} \varphi_1 = P_0^{(i)} V_{PQ} \frac{Q_0}{a} V_{QP}\varphi_0 / \left( \epsilon_1 - \epsilon_1^{(i)} \right), \quad i \neq 1. \quad (66)$$

Substituting into Eq. (65) we see that  $\epsilon_3$  is defined by the eigenvalue equation for the operator

$$\begin{aligned} V_2 \equiv P_2 V \frac{Q_0}{a} V \frac{Q_0}{a} V P_2 - \epsilon_1 P_2 V \frac{Q_0}{a^2} V P_2 \\ + \sum_{i \neq 1} P_2 V \frac{Q_0}{a} V \frac{P_0^{(i)}}{\epsilon_1 - \epsilon_1^{(i)}} V \frac{Q_0}{a} V P_2. \end{aligned} \quad (67)$$

Despite the presence of four factors in the potential, the last term is actually a third order term due to the denominators  $\epsilon_1 - \epsilon_1^{(i)}$ .

The procedure outlined above, which essentially embodies the Rayleigh–Schrödinger approach, can be pursued until the degeneracy is (if possible, see below Sect. “Symmetry and Degeneracy”) completely removed, after which the theory for the non-degenerate case can be used. Rather than detailing the calculations, we present an alternative iterative procedure due to Bloch [4] which allows a more systematic calculation of the corrections to the energy and the wave function.

### Bloch’s Method

In equations Eq. (51) and (52) we have seen that the energy corrections  $\Delta E$  and the projections onto  $W_0$  of the vectors  $\psi_k(\lambda)$  are eigenvalues and eigenvectors of an operator acting in  $W_0$ . This observation is not immediately useful since the operator depends on the unknown exact energy  $E(\lambda)$ . However, it is possible to produce an operator  $B(\lambda)$ , which can be calculated in terms of known quantities and has the property that, if  $E_k(\lambda)$ ,  $\psi_k(\lambda)$  are eigenvalues and eigenvectors of Eq. (10) such that  $E_k(0) = E_0$ , then

$$B(\lambda) P_0 \psi_k(\lambda) = \Delta E_k P_0 \psi_k(\lambda). \quad (68)$$

First of all, note that the vectors  $P_0 \psi_k(\lambda)$  are a basis for the subspace  $W_0$ . Indeed, it is implicit in the assumption that perturbation theory does work that the perturbing potential should produce only slight modifications of the unperturbed eigenvectors of the Hamiltonian, so that the vectors  $P_0 \psi_k(\lambda)$  are linearly independent (although not orthogonal). Since their number equals the dimension of  $W_0$ , they are a basis for this subspace.

Following [4], we define a  $\lambda$  dependent operator  $U$  in this way:

$$U P_0 \psi_k(\lambda) = \psi_k(\lambda); \quad U Q_0 = 0. \quad (69)$$

As a consequence we have

$$U = U P_0, \quad P_0 U = P_0, \quad (70)$$

$$U \psi_k(\lambda) = \psi_k(\lambda). \quad (71)$$

The former of Eq. (70) follows immediately from the definition of  $U$ . Hence  $P_0 U = P_0 U P_0$ , which implies the latter of Eq. (70). Equation (71) is verified by applying the former of Eq. (70) to  $\psi_k(\lambda)$ .

Let

$$B(\lambda) \equiv \lambda P_0 V U. \quad (72)$$

We verify that, if  $\Delta E_k \equiv E_k - E_0$ , then

$$B(\lambda) P_0 \psi_k(\lambda) = \Delta E_k P_0 \psi_k(\lambda). \quad (73)$$

Indeed, by Eq. (69) we have  $P_0 V U P_0 \psi_k(\lambda) = P_0 V \psi_k(\lambda)$ . Writing Eq. (10) as

$$(H_0 - E_0 + \lambda V) \psi_k(\lambda) = \Delta E_k \psi_k(\lambda)$$

and multiplying by  $P_0$  we find

$$\lambda P_0 V \psi_k(\lambda) = \Delta E_k P_0 \psi_k(\lambda), \quad (74)$$

hence Eq. (73) is satisfied.

A practical use of Eq. (73) requires an iterative definition of  $U$  in terms of known quantities. From Eqs. (69) and (70) we have

$$U = P_0 U + Q_0 U = P_0 + Q_0 U P_0. \quad (75)$$

We calculate the latter term of Eq. (75) on the vectors  $P_0 \psi_k(\lambda)$ . Since

$$(\lambda V - \Delta E_k) \psi_k = (E_0 - H_0) \psi_k,$$

recalling Eq. (69) we have

$$\begin{aligned} Q_0 U P_0 \psi_k(\lambda) &= Q_0 \psi_k(\lambda) \\ &= \frac{Q_0}{a} (\lambda V - \Delta E_k) \psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \Delta E_k \frac{Q_0}{a} U \psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \Delta E_k \frac{Q_0}{a} U P_0 \psi_k(\lambda). \end{aligned}$$

By Eq. (74)

$$\begin{aligned} Q_0 U P_0 \psi_k(\lambda) &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \lambda \frac{Q_0}{a} U P_0 V \psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \lambda \frac{Q_0}{a} U P_0 V U \psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} (V U - U V U) P_0 \psi_k(\lambda). \end{aligned}$$

As a consequence the desired iterative equation for  $U$  is

$$U = P_0 + \lambda \frac{Q_0}{a} (V U - U V U). \quad (76)$$

Equation (76) in turn allows an iterative definition of the operator  $B(\lambda)$  of Eq. (72) depending only on quantities which can be computed in terms of the known spectral representation of  $H_0$ . Knowing  $U$  through order  $n-1$  gives  $B^{[n]}(\lambda) \equiv \sum_{i=1}^n B^{(i)}(\lambda)$ , whose eigenvalues are the energy corrections through order  $n$ . In fact, if  $B = \sum_{i=1}^{\infty} B^{(i)}$  and  $P_0 \psi_k = \sum_{s=0}^{\infty} \lambda^s \varphi_s$ , the order  $r$  contribution to Eq. (73) is

$$\sum_{i=1}^r B^{(i)} \varphi_{r-i} = \sum_{i=1}^r \epsilon_i \varphi_{r-i}. \quad (77)$$

Defining  $P_0 \psi_k^{[n]} \equiv \sum_0^n \lambda^r \varphi_r \equiv \varphi^{[n]}$ ,  $\Delta E^{[n]} \equiv \sum_1^n \lambda^r \epsilon_r$ , we see that the sum of Eq. (77) for values of  $r$  through  $n$  gives

$$B^{[n]} \varphi^{[n]} = \Delta E^{[n]} \varphi^{[n]} + O(\lambda^{n+1}). \quad (78)$$

Once  $P_0 \psi_k^{[n]}(\lambda)$  has been found, Eq. (69) gives the component of  $\psi_k(\lambda)$  in  $W_0^\perp$  through order  $n+1$ . As an example, for  $n=3$  we have

$$\begin{aligned} U^{[2]} &= P_0 + \lambda \frac{Q_0}{a} V P_0 + \lambda^2 \frac{Q_0}{a} V \frac{Q_0}{a} V P_0 \\ &\quad - \lambda^2 \frac{Q_0}{a^2} V P_0 V P_0, \quad (79) \end{aligned}$$

$$\begin{aligned} B^{[3]} &= \lambda P_0 V P_0 + \lambda^2 P_0 V \frac{Q_0}{a} V P_0 + \lambda^3 P_0 V \frac{Q_0}{a} V \frac{Q_0}{a} V P_0 \\ &\quad - \lambda^3 P_0 V \frac{Q_0}{a^2} V P_0 V P_0. \quad (80) \end{aligned}$$

If  $W_0$  is one dimensional, Eq. (80) gives for  $\lambda \epsilon_1 + \lambda^2 \epsilon_2 + \lambda^3 \epsilon_3$  the same result as Eqs. (27), (28) and (29).

The main difference between the RS perturbation theory and Bloch's method is that within the former the energy corrections through order  $n$  are calculated by means of a sequential computation starting from  $\epsilon_1$ , with the consequence that at each step the dimension of the matrix to be diagonalized is smaller. Conversely, within Bloch's method one has to diagonalize the matrix  $B^{[n]}(\lambda)$ , which has the dimension of  $W_0$ . However, as noted above, for  $n > 1$  the eigenvalues of  $B^{[n]}(\lambda)$  are different from  $\lambda \epsilon_1 + \lambda^2 \epsilon_2 + \dots + \lambda^n \epsilon_n$  by terms of order at least  $n+1$ . Similarly, the eigenvectors of Eq. (78) differ from the component in  $W_0$  of  $\psi^{[n]} = \psi_0 + \lambda \psi^{(1)} + \dots + \lambda^n \psi^{(n)}$  by terms of order larger than  $n$ .

It is instructive to reconsider the calculation of  $\epsilon_2$  and  $\epsilon_3$  in the light of Bloch's method. If  $P_0 = P_1 + P'_1$ , then  $P_0 V P_0 = \epsilon_1 P_1 + P'_1 V P'_1$  and

$$\begin{aligned} B^{[2]}(\lambda) &= \lambda \epsilon_1 P_1 + \lambda P'_1 V P'_1 + \lambda^2 P_1 V \frac{Q_0}{a} V P_1 \\ &\quad + \lambda^2 P'_1 V \frac{Q_0}{a} V P'_1 + \lambda^2 P_1 V \frac{Q_0}{a} V P'_1 + \lambda^2 P'_1 V \frac{Q_0}{a} V P_1. \end{aligned}$$

The last two terms represent off-diagonal blocks which can be omitted for the calculation of  $\lambda \epsilon_1 + \lambda^2 \epsilon_2$ , since the lowest order contribution to the eigenvalues of a matrix  $X$  from the off-diagonal terms  $X_{ij}$  is  $|X_{ij}|^2/(X_{ii} - X_{jj})$ . For a second order expansion as  $B^{[2]}$  this yields third order contributions of the type

$$\lambda^3 P_1 V \frac{Q_0}{a} V \frac{P'_1}{\epsilon_1 - \epsilon'_1} V \frac{Q_0}{a} V P_1.$$

These are just the contributions to  $\epsilon_3$  which we met in the RS approach: the expression of  $V_2$  given in Eq. (67) combines the block-diagonal term of order 3 with the off-diagonal terms of order 2 giving a third order contribution.

### The Quasi-Degenerate Case

There are cases, in both atomic and molecular physics, where the energy levels of  $H_0$  present a multiplet structure: the energy levels are grouped into “multiplets” whose separation  $\Delta E$  is large compared to the energy separation  $\delta E$  between the levels belonging to the same multiplet. For instance, in atomic physics this is the case of the fine structure (due to the so called spin-orbit interaction) or of the hyperfine structure (due to the interaction of the nuclear magnetic moment with the electrons); in molecular physics typically this is the case of the rotational levels associated with the different and widely separated vibrational levels.

If a perturbation  $V$  is such that its matrix elements between levels of the same multiplet are comparable to  $\delta E$ , while being small with respect to  $\Delta E$ , then naive perturbation theory fails because of the small energy denominators pertaining to levels belonging to the same multiplet. To solve this problem, named the problem of quasi-degenerate levels, once again we can exploit the arbitrariness in the way of splitting the Hamiltonian  $H$  into an unperturbed Hamiltonian and a perturbation. Let

$$E_0^{(1)} \equiv E_0 + \delta E^{(1)}, \quad E_0^{(2)} \equiv E_0 + \delta E^{(2)}, \quad \dots, \\ E_0^{(n)} \equiv E_0 + \delta E^{(n)},$$

be the unperturbed energies within a multiplet, with  $E_0$  any value close to the  $E_0^{(i)}$ 's (for instance their mean value), and  $P_0^{(i)}$  the projections onto the corresponding eigenspaces. Let

$$H_0^0 \equiv H_0 - \sum_i \delta E^{(i)} P_0^{(i)}, \quad \tilde{V} \equiv \lambda V + \sum_i \delta E^{(i)} P_0^{(i)},$$

so that

$$H = H_0^0 + \tilde{V}. \quad (81)$$

We consider  $H_0^0$  as the unperturbed Hamiltonian and  $\tilde{V}$  as the perturbation. From the physical point of view this procedure, if applied to all multiplets, is just the inclusion into the perturbation of those terms of  $H_0$  that are responsible for the multiplet structure. With the splitting of the Hamiltonian as in Eq. (81) we can apply the methods of degenerate perturbation theory. The most efficient

of these techniques is Bloch's method, which yields a simple prescription for the calculation of the corrections of any order. If for example we are content with the lowest order, we must diagonalize the matrix  $P_0 \tilde{V} P_0$ , or equivalently  $P_0 H P_0$ , that is the energies through first order are the eigenvalues of the equation

$$P_0 H P_0 \psi = E P_0 \psi, \quad (82)$$

where  $P_0 = \sum_i P_0^{(i)}$  is the projection onto  $W_0$ , the eigenspace of  $H_0^0$  corresponding to the eigenvalue  $E_0$ . These eigenvalues are algebraic functions of  $\lambda$ , and no finite order approximation is meaningful, since all terms can be numerically of the same order, due to the occurrence of small denominators  $(\delta E^{(i)} - \delta E^{(j)})^n$ .

### The Brillouin-Wigner Method

Equations (52) and (55) yield an alternative approach to the calculation of the energy shift  $\Delta E$  due to a perturbation to a non-degenerate energy level  $E_0$ , the so called Brillouin-Wigner method [9,22,52]. In this case  $W_0$ , the space spanned by the unperturbed eigenvector  $\psi_0$ , is one-dimensional. The correction  $\Delta E$  obeys the equation

$$\Delta E = (\psi_0, A(E) \psi_0), \quad (83)$$

where the operator  $A(E)$  is defined in Eq. (52).

Substituting into the expression of  $A$  the expansion Eq. (55) for  $(E - H_{QQ})^{-1}$  and noting that, if  $\{E_k\}$  is the spectrum of  $H_0$ ,

$$(\psi_0, V_{PQ}(E - Q_0 H_0 Q_0)^{-1} V_{QP} \psi_0) = \sum_{k \neq 0} \frac{|V_{0k}|^2}{E - E_k},$$

$$(\psi_0, V_{PQ}(E - Q_0 H_0 Q_0)^{-1} V_{QQ}(E - Q_0 H_0 Q_0)^{-1} V_{QP} \psi_0) = \sum_{k, h \neq 0} V_{0k}(E - E_k)^{-1} V_{kh}(E - E_h)^{-1} V_{h0}$$

and so on, we find the following implicit expression for the exact energy  $E$ :

$$E = E_0 + \lambda(\psi_0, V \psi_0) + \lambda^2 \sum_{k \neq 0} \frac{|V_{0k}|^2}{E - E_k} \\ + \lambda^3 \sum_{k, h \neq 0} \frac{V_{0k}}{E - E_k} \frac{V_{kh}}{E - E_h} V_{h0} + \dots \quad (84)$$

Consistently with the assumption that perturbation theory does work, the denominators in Eq. (84) are non-vanishing. The equation can be solved by arresting the expansion to a given power  $n$  in the potential and searching

a solution iteratively starting with  $E = E_0$ . However, the result differs from the energy  $E^{[n]} = E_0 + \lambda \epsilon_1 + \lambda^2 \epsilon_2 + \dots + \lambda^n \epsilon_n$ , calculated by means of the RS perturbation theory, by terms of order  $n + 1$  in the potential. The result of the RS perturbation theory can be recovered from the Brillouin–Wigner approach by substituting in the denominators  $E = E_0 + \lambda \epsilon_1 + \lambda^2 \epsilon_2 + \dots + \lambda^n \epsilon_n$ , expanding the denominators in powers of  $\lambda^k \epsilon_k / E_0$  and equating terms of equal orders in both sides of Eq. (84).

As for the perturbed wave function, if the intermediate normalization is used, by Eq. (49) we have:

$$\psi = \psi_0 + Q_0 \psi = \psi_0 + \lambda(E - H_{QQ})^{-1} V_{QP} \psi_0. \quad (85)$$

Again, using the expansion Eq. (55) we find

$$\begin{aligned} \psi = \psi_0 + \lambda \sum_{k \neq 0} \psi_k \frac{V_{k0}}{E - E_k} \\ + \lambda^2 \sum_{k, h \neq 0} \psi_k \frac{V_{kh}}{E - E_k} \frac{V_{h0}}{E - E_h} + \dots \end{aligned} \quad (86)$$

As for the energy, if we arrest this expression to order  $n$  and substitute for  $E$  the value calculated by using Eq. (84), the result will differ from the one of Rayleigh–Schrödinger perturbation theory by terms of order  $n + 1$ .

A major drawback of the Brillouin–Wigner method is its lack of size-consistency: for a system consisting of non-interacting subsystems, the perturbative correction to the energy of the total system is not the sum of the perturbative corrections to the energies of the separate subsystems through any finite order. This is best illustrated by the simple case of two systems  $a, b$  with unperturbed eigenvectors, energies and interactions  $\psi_0^a, E_0^a, \lambda V^a$  and  $\psi_0^b, E_0^b, \lambda V^b$  respectively. If for example the expansion Eq. (84) is arrested at order 2, by noting that the matrix elements  $V_{0,ij}$  between the unperturbed state and the states  $\psi_i^a \psi_j^b$  are

$$\begin{aligned} V_{0,ij} &\equiv \left( \psi_0^a \psi_0^b, (V^a + V^b) \psi_i^a \psi_j^b \right) \\ &= \left( \psi_0^a, V^a \psi_i^a \right) \delta_{0j} + \left( \psi_0^b, V^b \psi_j^b \right) \delta_{0i}, \end{aligned}$$

for the second order equation defining  $E$  we find

$$\begin{aligned} E = E_0^a + E_0^b + \lambda \epsilon_1^a + \lambda \epsilon_1^b + \lambda^2 \sum_j \frac{|V_{0j}^b|^2}{E - E_0^a - E_j^b} \\ + \lambda^2 \sum_i \frac{|V_{0i}^a|^2}{E - E_0^b - E_i^a}. \end{aligned} \quad (87)$$

On the other hand, for the energy of each system at second order we find

$$\begin{aligned} E^a &= E_0^a + \lambda \epsilon_1^a + \lambda^2 \sum_i \frac{|V_{0i}^a|^2}{E_a - E_i^a}; \\ E^b &= E_0^b + \lambda \epsilon_1^b + \lambda^2 \sum_j \frac{|V_{0j}^b|^2}{E_b - E_j^b}. \end{aligned} \quad (88)$$

It is apparent that the sum of the expression reported in Eq. (88) does not equal the expression of the energy reported in Eq. (87). This pathology is absent in the RS perturbation theory, where for non-interacting systems  $E(\lambda) = E^a(\lambda) + E^b(\lambda)$ , hence, for any  $j, \epsilon_j = (1/j!) D^j E(\lambda)|_{\lambda=0} = \epsilon_j^a + \epsilon_j^b$ .

## Symmetry and Degeneracy

In Sect. “[Perturbation of Point Spectra: Degenerate Case](#)” we applied perturbation theory to the case of degenerate eigenvalues with special emphasis on the problem of the removal of the degeneracy at a suitable order of perturbation theory. The main problem is to know in advance whether the degeneracy can be removed completely, or a residual degeneracy is to be expected. The answer is given by group theory [21,51,53].

The very existence of degenerate eigenvalues of a Hamiltonian  $H$  is intimately connected with the symmetry properties of this operator. Generally speaking, a group  $G$  is a symmetry group for a physical system if there exists an associated set  $\{T(g)\}$  of transformations in the Hilbert space of the system such that  $|(T(g)\varphi, T(g)\psi)|^2 = |(\varphi, \psi)|^2, g \in G$  [53]. It is proven that the operators  $T(g)$  must be either unitary or antiunitary [2,53]. We will consider the most common case that they are unitary and can be chosen in such a way that

$$T(g_1)T(g_2) = T(g_1g_2), \quad g_1, g_2 \in G, \quad (89)$$

so that the operators  $\{T(g)\}$  are a representation of  $G$ .

A system described by a Hamiltonian  $H$  is said to be invariant under the group  $G$  if the time evolution operator commutes with  $T(g)$ . Under fairly wide hypotheses this implies

$$[H, T(g)] = 0, \quad g \in G. \quad (90)$$

A consequence is that, for any  $g \in G$ ,

$$H\psi = E\psi \Rightarrow HT(g)\psi = ET(g)\psi, \quad (91)$$

that is the restrictions  $T(g)|_W$  of the operators  $T(g)$  to the space  $W$  corresponding to a given energy  $E$  are a representation of  $G$ . Given an orthonormal basis  $\{\psi_i\}$  in  $W$ , we

have

$$T(g)\psi_i = \sum_j t_{ji}(g)\psi_j \quad (92)$$

and the vectors  $\psi_i$  are said to transform according to the representation of  $G$  described by the matrices  $t_{ji}$ .

This representation, apart from the occurrence of the so called accidental degeneracy (which in most cases actually is a consequence of the invariance of the Hamiltonian under additional transformations) is irreducible: no subspace of  $W$  is invariant under all the transformations of the group. As a consequence, knowing the dimensions  $d_j$  of the irreducible representations of  $G$  allows to predict the possible degree of degeneracy of a given energy level, since the dimension of  $W$  must be equal to one of the numbers  $d_j$ . If the group of invariance is Abelian, all the irreducible representations are one dimensional, and degeneracy can only be accidental.

Two irreducible representations are equivalent if there are bases which transform with the same matrix  $t_{ji}(g)$ . Otherwise they are inequivalent. The following orthogonality theorems hold. If  $a$  and  $b$  are inequivalent representations and  $\psi_i^{(a)}, \varphi_j^{(b)}$  transform according these representations, then

$$\left( \psi_i^{(a)}, \varphi_j^{(b)} \right) = 0 \quad (93)$$

while, if  $a_r$  and  $a_s$  are equivalent, for the basis vectors  $\psi_i^{(a_r)}, \varphi_j^{(a_s)}$  we have

$$\left( \psi_i^{(a_r)}, \varphi_j^{(a_s)} \right) = K_{rs}^{(a)} \delta_{ij}. \quad (94)$$

Moreover, if  $A_{rs}$  is a matrix which commutes with all the matrices  $t_{ij}^{(a)}$  of an irreducible representation  $b$ , then  $A_{rs} = a\delta_{rs}$  (Schur's lemma).

### Symmetry and Perturbation Theory

If  $H = H_0 + \lambda V$ , let  $G_0$  be the group under which  $H_0$  is invariant. Although it is not the commonest case, we start with assuming that also the perturbation  $V$  commutes with  $T(g)$  for any element  $g$  of  $G_0$ . As a rule  $W_0$ , the space of eigenvectors of  $H_0$  with energy  $E_0$ , hosts an irreducible representation  $T(g)$  of  $G_0$ . In this case the degeneracy cannot be removed at any order of perturbation theory. While this follows from general principles (for any value of  $\lambda$ ,  $\psi(\lambda)$  and  $T(g)\psi(\lambda)$  are eigenvectors of  $H(\lambda)$ , and by continuity the eigenspace  $W_\lambda$  will have the same dimension as  $W_0$ ), it is interesting to understand how the symmetry properties affect the mechanism of perturbation theory.

If  $\{\psi_i^0\}$  is a basis of  $W_0$  transforming according to an irreducible representation  $a$  of  $G_0$ , then the matrix  $V_{ij} = (\psi_i^0, V\psi_j^0)$  commutes with all the matrices  $t_{ji}^{(a)}(g)$  and, according to Schur's lemma,  $(\psi_i^0, V\psi_j^0) = v\delta_{ij} = \epsilon_1\delta_{ij}$ . No splitting occurs at the level of first order perturbation theory, neither can it occur at any higher order. Indeed, when  $V$  commutes with the operators  $T(g)$ , then Bloch's operator  $U$ , and consequently the operator  $B(\lambda)$  of Eq. (72), both commute with the  $T(g)$ 's too. Again by Schur's lemma, the operator  $B(\lambda)$  is a multiple of the identity. At any order of perturbation the degeneracy of the level is not removed.

In most of the cases, however, the perturbation  $V$  does not commute with all the operators  $T(g)$ . The set

$$G = \{g : g \in G_0, [T(g), V] = 0\}$$

is a subgroup  $G$  of  $G_0$  and the group of invariance for the Hamiltonian  $H$  is reduced to  $G$ .  $W_0$  generally contains  $G$ -irreducible subspaces  $W_i$ ,  $1 \leq i \leq n$ : the operators  $T(g)|_{W_0}$  are a reducible representation of  $G$ . The decomposition into irreducible representations of  $G$  is unique up to equivalence.

The crucial information we gain from group theory is the following: the number of energy levels which the energy  $E_0$  is split into is the number of irreducible representations of  $G$  which the representation of  $G_0$  in  $W_0$  is split into. The degrees of degeneracy are the dimensions of these representations. What is relevant is that we only need to study the eigenspace  $W_0$  of  $H_0$ , which is known by hypothesis.

In fact, let  $W(\lambda)$  be the space spanned by the eigenvectors  $\psi_k(\lambda)$  of  $H(\lambda)$  such that  $\psi_k(0) \in W_0$ .  $W(\lambda)$  is invariant under the operators  $T(g)$ ,  $g \in G$ , since Bloch's operator  $U$  commutes with the operators  $T(g)$ ,  $g \in G$ .  $W(\lambda)$  can be decomposed into  $G$ -irreducible subspaces  $W_k(\lambda)$ , and in each of them by Schur's lemma the Hamiltonian  $H(\lambda)$  is represented by a matrix  $E_k(\lambda)I_{W_k(\lambda)}$ . The projections  $P_0 W_k(\lambda)$  span the space  $W_0$  and transform with the same representation of  $G$  as  $W_k(\lambda)$ , since  $P_0$  commutes with  $T(g)$  for any  $g$  in  $G_0$ , hence for any  $g$  in  $G$ . Thus, the space  $W_0$  hosts as many irreducible representations of  $G$  as  $W(\lambda)$ . Assuming that the eigenvalues  $E_k(\lambda)$  are different from one another, we see that the decomposition of the representation of  $G_0$  in  $W_0$  into irreducible representations of  $G$  determines the number and the degeneracy of the eigenvalues of  $H(\lambda)$  such that the corresponding eigenvectors  $\psi(\lambda)$  are in  $W_0$  for  $\lambda = 0$ . The possibility that some of the  $E_k(\lambda)$  are equal will be touched upon in the next subsection.

Examples where the above mechanism is at work are common in atomic physics. When an atom, whose unper-



turbed Hamiltonian  $H_0$  is invariant under  $O(3)$ , is subjected to a constant electric field  $\vec{E} = E\hat{z}$  (Stark effect), the invariance group  $G$  of its Hamiltonian is reduced to the rotations about the  $z$  axis ( $SO(2)$ ) and the reflections with respect to planes containing the  $z$  axis. The irreducible representations of this group have dimension at most 2, and the  $G$ -irreducible subspaces of  $W_0$  (the space generated by the eigenvectors  $\psi_{E_0lm}$  of  $H_0$  corresponding to the energy  $E_0$ ) are generated by  $\psi_{E_0l0}$  (one dimensional representation) and  $\psi_{E_0l\pm m}$  (two dimensional representations). Hence, the level  $E_0$  is split into  $l + 1$  levels, the states with  $m$  and  $-m$  remaining degenerate since reflections transform a vector with a given  $m$  into the vector with opposite  $m$ . Instead, if the atom is subjected to a constant magnetic field  $\vec{B} = B\hat{z}$  (Zeeman effect), the surviving invariance group  $G$  consists of  $SO(2)$  plus the reflections with respect to planes  $z = z_0$ .  $G$  being Abelian, the degeneracy is completely removed and this occurs at the first order of perturbation theory.

In the rather special case that  $W_0$  contains subspaces transforming according to inequivalent representations of  $G_0$ , also a  $G_0$ -invariant perturbation  $V$  can separate in energy the states belonging to inequivalent representations. For example, the spectrum of alkali atoms can be calculated by considering in a first approximation an electron in the field of the unit charged atomic rest, which is treated as pointlike. In this problem the obvious invariance group of the Hamiltonian of the optical electron is  $O(3)$ , the group of rotations and reflections, and the space  $W_0$  corresponding to the principal quantum number  $n > 1$  contains  $n$  inequivalent irreducible representations which are labeled by the angular momentum  $l \leq n - 1$ . When the finite dimension of the atomic rest is taken into account as a perturbation, its invariance under  $O(3)$  splits the levels with given  $n$  and different  $l$  into  $n$  sublevels. A more careful consideration, however, shows that also the Lenz vector commutes with the unperturbed Hamiltonian [5], and that the space  $W_0$  is irreducible under a larger group, the group  $SO(4)$  [18], which is generated by the angular momentum and the Lenz vector. As a consequence the  $l$  degeneracy is by no means accidental: a space irreducible under a given group can turn out to be reducible with respect to one of its subgroups.

Group theory is a valuable tool in degenerate perturbation theory to search the correct vectors  $\psi_k(0)$  which make the operator  $P_0 V P_0$  diagonal. In fact, let  $\psi_i^{(a)}$  be vectors which reduce the representation  $T$  of  $G$  in  $W_0$  into its irreducible components  $T^{(a)}$ . The vectors  $\psi_i^{(a)}$  and  $V\psi_i^{(a)}$  transform according to the same irreducible representation  $T^{(a)}$ . Hence, by Eqs. (93) and (94) we find that the  $P_0 V P_0$  is a diagonal block matrix with respect to inequiva-

lent representations:

$$\left( \psi_i^{(a)}, V \psi_j^{(b)} \right) = K_{rs}^{(a)} \delta_{ij} \delta_{ab}, \quad (95)$$

with  $\delta_{ab} = 1$  if representations  $a$  and  $b$  are equivalent,  $\delta_{ab} = 0$  otherwise. The matrices  $K_{rs}^{(a)}$  are generally much smaller than the full matrix of the potential. Thus, the operation of diagonalizing  $V$  is made easier by finding the  $G$ -irreducible subspaces  $W_a$ . Conversely, the reduction of an irreducible representation of a group  $G_0$  in a space  $W_0$  into irreducible representations of a subgroup  $G$  can be achieved by the following trick: find an operator  $V$  whose symmetry group is just  $G$  and interpret  $W_0$  as the degeneracy eigenspace of a Hamiltonian  $H_0$ . The  $G$ -irreducible subspaces of  $W_0$  are the eigenspaces of  $P_0 V P_0$ .

### Level Crossing

As shown in the foregoing section, the existence of a non-Abelian group of symmetry for the Hamiltonian entails the existence of degenerate eigenvalues. The problem naturally arises as to whether there are cases when, on the contrary, the degeneracy is truly “accidental”, that is it cannot be traced back to symmetry properties. The problem was discussed by J. Von Neumann and E.P. Wigner [49], who showed that for a generic  $n \times n$  Hermitian matrix depending on real parameters  $\lambda_1, \lambda_2, \dots$ , three real values of the parameters have to be adjusted in order to have the collapse of two eigenvalues (level crossing).

When passing to infinite dimension, arguments valid for finite dimensional matrices might fail. Moreover, often the Hamiltonian is not sufficiently “generic” so that level crossing may occur. As a consequence, we look for necessary conditions in order that, given the Hamiltonian  $H(\lambda) = H_0 + \lambda V$ , two eigenvalues collapse for some (real) value  $\bar{\lambda}$  of the parameter  $\lambda$ :  $E_1(\bar{\lambda}) = E_2(\bar{\lambda}) \equiv \bar{E}$ . In this case, if  $\psi_1(\bar{\lambda})$  and  $\psi_2(\bar{\lambda})$  are any two orthonormal eigenvectors of  $H(\bar{\lambda}) = H_0 + \bar{\lambda} V$  belonging to the eigenvalue  $\bar{E}$ , the matrix

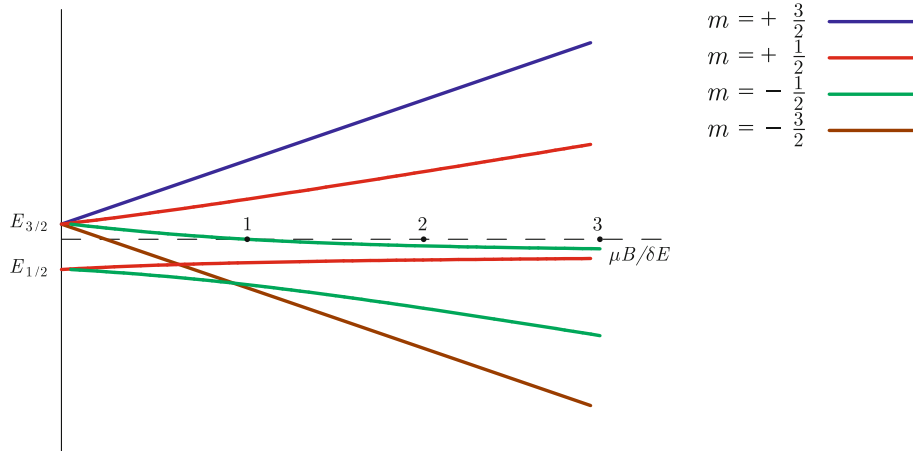
$$H_{ij}(\bar{\lambda}) \equiv (\psi_i(\bar{\lambda}), (H_0 + \bar{\lambda} V) \psi_j(\bar{\lambda})) \quad , \quad i, j = 1, 2$$

must be a multiple of the identity:

$$H_{11}(\bar{\lambda}) = H_{22}(\bar{\lambda}) \quad , \quad (96)$$

$$H_{12}(\bar{\lambda}) = 0 \quad . \quad (97)$$

Equations (96) and (97) are three real equations for the unknown  $\bar{\lambda}$ ; hence, except for special cases, level crossing cannot occur.



**Perturbation Theory in Quantum Mechanics, Figure 2**

The effect of a magnetic field on the doublet  $2p_{1/2}, 2p_{3/2}$  of the lithium whose degeneracies, in the absence of the magnetic field, are respectively 2 and 4.  $\mu$  is the magnetic moment of the electron and  $\mu B / \delta E = 1$  for  $B \approx 1.4$  T

The condition expressed by Eq. (97) is satisfied if the states corresponding to the eigenvalues  $E_1(\lambda)$  and  $E_2(\lambda)$  possess different symmetry properties, that is if they belong to inequivalent representations of the invariance group of the Hamiltonian or, equivalently, if they are eigenvectors with different eigenvalues of an operator which for any  $\lambda$  commutes with the Hamiltonian  $H(\lambda)$  (hence it commutes with both  $H_0$  and  $V$ ). In this case  $H_{12} = 0$  and the occurrence of level crossing depends on whether Eq. (96) has a real solution. This explains the statement that level crossing can occur only for states with different symmetry, while states of equal symmetry repel each other. Indeed, if Eq. (97) is not satisfied, the behavior of two close eigenvalues as functions of  $\lambda$  is illustrated in Fig. 1 (Sect. “Presentation of the Problem and an Example”).

Figure 2 illustrates the behavior of the quasi-degenerate energy levels  $2p_{1/2}, 2p_{3/2}$  of the lithium atom in the presence of an external magnetic field  $\vec{B}$ . In the absence of the magnetic field they are split by the spin-orbit interaction, with a separation  $\delta E \equiv E_{3/2} - E_{1/2} = 0.4 \times 10^{-4}$  eV, to be compared with the separation in excess of 1 eV from the adjacent  $2s$  and  $3s$  levels. This justifies treating the effect of the magnetic field by means of the first order perturbation theory for quasi-degenerate levels.

When the magnetic field is present, the residual symmetry is the (Abelian) group of rotations about the direction of  $\vec{B}$ . Hence, the Hamiltonian commutes with the component of the angular momentum along the direction of  $\vec{B}$ , whose eigenvalues are denoted with  $m$ . In Fig. 2 the energies of states with equal symmetry, that is with the same value of  $m$ , are depicted with the same color. No crossing occurs between states with equal  $m$ , while

the level with  $m = -3/2$  does cross both the levels with  $m = 1/2$  and with  $m = -1/2$  which the  $2p_{1/2}$  level is split into.

### Problems with the Perturbation Series

So far we have assumed that all the power expansions appearing in the calculations were converging for  $|\lambda| \leq 1$ , that is we assumed analyticity in  $\lambda$  of  $E(\lambda)$ . Actually, it is only for rather special cases that analyticity can be proved. For most of the cases of physical interest, even if the terms of the perturbation series can be shown to exist, the series does not converge, or, when it converges, the limit is not  $E(\lambda)$ . In spite of this, special techniques have been devised to extract a good approximation to  $E(\lambda)$  from the (generally few) terms of the perturbation series which can be computed. We will outline the main results existing in the field, without delving into mathematical details, for which we refer the reader to the books of Kato [25] and Reed–Simon [34] and the references therein.

The most favorable case is that of the so called regular perturbations [36,37,38,39], where the perturbation series does converge to  $E(\lambda)$ . More precisely, if  $E_0$  is a nondegenerate eigenvalue of  $H_0$ , for  $\lambda$  in a suitable neighborhood of  $\lambda = 0$  the Hamiltonian  $H = H_0 + \lambda V$  has a nondegenerate eigenvalue  $E(\lambda)$  which is analytic in  $\lambda$  and equals  $E_0$  for  $\lambda = 0$ . The same property holds for the eigenvector  $\psi(\lambda)$ . A sufficient condition for this property to hold is expressed by the Kato–Rellich theorem [26,36,37,38,39], which essentially states that if the perturbation  $V$  is  $H_0$ -bounded, in the sense that constants  $a, b$  exist such that

$$\|V\psi\| \leq a\|H_0\psi\| + b\|\psi\| \quad (98)$$

for any  $\psi$  in the domain of  $V$  (which must include the domain of  $H_0$ ) then the perturbation is regular. A lower bound to the radius  $r$  such that the perturbation series converges to the eigenvalue  $E(\lambda)$  for  $|\lambda| < r$  can be given in terms of the parameters  $a$ ,  $b$  appearing in Eq. (98) and the distance  $\delta$  of the eigenvalue  $E_0$  from the rest of the spectrum of  $H_0$ . We have

$$r = \left[ a + \frac{2}{\delta} \left[ b + a \left( |E_0| + \frac{\delta}{2} \right) \right] \right]^{-1}. \quad (99)$$

It must be stressed, however, that the constants  $a$  and  $b$  are not uniquely determined by  $V$  and  $H_0$ . If the perturbation  $V$  is bounded ( $a = 0$ ,  $b = \|V\|$ ) condition Eq. (99) reads  $r = \delta/(2\|V\|)$ , which implies that the perturbation series for  $H = H_0 + V$  with  $V$  bounded converges if  $\|V\| < \delta/2$  (Kato bound [26]). The analysis of the two-level system (Sect. “Presentation of the Problem and an Example”) shows that the figure 1/2 cannot be improved. Still, Kato bound is only a lower bound to  $r$ .

A similar statement holds for degenerate eigenvalues [34]: if  $E_0$  has multiplicity  $m$  there are  $m$  single valued analytic functions  $E_k(\lambda)$ ,  $k = 1, \dots, m$  such that  $E_k(0) = E_0$  and, for  $\lambda$  in a neighborhood of 0,  $E_k(\lambda)$  are eigenvalues of  $H(\lambda) = H_0 + \lambda V$ . Some of the functions  $E_k(\lambda)$  may be coincident, and in a neighborhood of  $E_0$  there are no other eigenvalues of  $H(\lambda)$ .

Regular perturbations are in fact exceedingly rare, a notable case being that of helium-like atoms [47]. Actually, there are cases where, although on physical grounds  $H_0 + \lambda V$  does possess bound states, the relationship between  $E(\lambda)$  and the RS expansion is far more complicated than for regular perturbations. As pointed out by Kramers [27], with an argument similar to an observation by Dyson [12] for quantum electrodynamics, the quartic anharmonic oscillator with Hamiltonian

$$H = H_0 + \lambda V \equiv \frac{p^2}{2m} + \frac{m\omega^2 x^2}{2} + \lambda \frac{m^2 \omega^3}{\hbar} x^4 \quad (100)$$

is such an example. In fact, on the one hand bound states exist only for  $\lambda \geq 0$ ; on the other hand, if a power series converges for  $\lambda > 0$ , then the series should converge also for negative values of  $\lambda$ . But for  $\lambda < 0$  no bound state exists. Still worse, by estimating the coefficients of the RS expansion it has been proved that the series has vanishing radius of convergence [3].

In spite of this negative result, in this case it has been proved [46] that the perturbation series is an asymptotic series. This means that, for each  $n$ , if  $\sum_0^n \epsilon_k \lambda^k$  is the sum through order  $n$  of the perturbation series, then

$$\lim_{\lambda \rightarrow 0} \frac{\sum_0^n \epsilon_k \lambda^k - E(\lambda)}{\lambda^n} = 0. \quad (101)$$

We recall the difference between an asymptotic and an absolutely converging series, such as occurs with regular perturbations. For the latter one, given any  $\lambda$  in the convergence range of the series, the distance  $|\sum_0^n \epsilon_k \lambda^k - E(\lambda)|$  can be made arbitrarily small provided  $n$  is sufficiently large (so that a converging series is also an asymptotic series). On the contrary, for an asymptotic series  $|\sum_0^n \epsilon_k \lambda^k - E(\lambda)|$  is arbitrarily small only if  $\lambda$  is sufficiently near 0, but for a definite value of  $\lambda$  the quantity  $|\sum_0^n \epsilon_k \lambda^k - E(\lambda)|$  might decrease to a minimum, attained for some value  $N$ , and then it could start to oscillate for  $n > N$  (this is indeed the case for the anharmonic oscillator). As a consequence, for asymptotic series it is not expedient to push the calculation of the terms of the series beyond the limit where wild oscillations set in.

Any  $C^\infty$  function has an asymptotic series, as can be seen by inspection of the Taylor’s formula with a remainder (see Eq. (42)). By this means Krieger [28] argued that, if  $\epsilon_k$  (or equivalently the  $k$ th derivative of  $E(\lambda)$ ) exists for any  $k$ , the RS series is asymptotic. However, generally there is not a range where the series converges to  $E(\lambda)$ , that is  $E(\lambda)$  is not analytic. An asymptotic series may fail to converge at all for  $\lambda \neq 0$ , as noted for the anharmonic oscillator. The asymptotic series of a function, if it exists, is unique, but the converse is not true. For example, for the  $C^\infty$  function defined for real  $x$  as  $f(x) = \exp(-1/x^2)$  if  $x \neq 0$ ,  $f(0) = 0$ , the asymptotic series vanishes. There are also cases when the perturbation series is asymptotic for  $\arg \lambda$  lying in a range  $[\alpha, \beta]$ . This occurs for example for the generalized anharmonic oscillator with perturbation  $V \propto \lambda x^{2n}$ . It has been proved that its perturbation series is asymptotic for  $|\arg \lambda| \leq \theta < \pi$  [46] (note that the domain does not include negative values of  $\lambda$ ). The result was later extended to multidimensional anharmonic oscillators [19]. General theorems stating sufficient hypotheses for the perturbation series to be asymptotic can be found in the literature. As a rule, however, they do not cover most of the cases of physical interest.

Even in the felicitous case when the perturbation series is asymptotic, it is only known that a partial sum approaches  $E(\lambda)$  as much as desired provided  $\lambda$  is sufficiently small. This is not of much help to the practicing scientist, who generally is confronted with a definite value of the parameter  $\lambda$ , which can always be considered  $\lambda = 1$  by an appropriate rescaling of the potential  $V$ . Recalling that different functions can have the same asymptotic series, it seems hopeless to try to recover the function  $E(\lambda)$  from its asymptotic series, but this is possible for the so called strong asymptotic series. A function  $E(\lambda)$  analytic in a sectorial region ( $0 < |\lambda| < B$ ,  $|\arg \lambda| < \pi/2 + \delta$ ) is said to have strong asymptotic series  $\sum_0^\infty a_k \lambda^k$  if for all  $\lambda$

in the sector

$$\left| E(\lambda) - \sum_0^n a_k \lambda^k \right| \leq C \sigma^{n+1} |\lambda|^{n+1} (n+1)! \quad (102)$$

for some constants  $C, \sigma$ . For strong asymptotic series it is proved that the function  $E(\lambda)$  is uniquely determined by the series. Conditions that ensure that the RS series is a strong asymptotic series have been given [34].

The problem of actually recovering the function  $E(\lambda)$  from its asymptotic series can be tackled by several methods. The most widely used procedure is the Borel summation method [6], which amounts to what follows. Given the strongly asymptotic series  $\sum_0^\infty a_k \lambda^k$ , one considers the series  $F(\lambda) \equiv \sum_0^\infty (a_k/k!) \lambda^k$ . This is known as the Borel transform of the initial series, which, by the hypothesis of strong asymptotic convergence, can be proved to have a non-vanishing radius of convergence and to possess an analytic continuation to the positive real axis. Then the function  $E(\lambda)$  is given by

$$E(\lambda) = \int_0^\infty F(\lambda x) \exp(-x) dx. \quad (103)$$

The above statement is Watson's theorem [50]. Roughly speaking, it yields the function  $E(\lambda)$  as if the following exchange of the series with the integral were allowed:

$$\begin{aligned} E(\lambda) &\sim \sum_0^\infty a_k \lambda^k = \sum_0^\infty \frac{a_k}{k!} \int_0^\infty \exp(-x) x^k dx \lambda^k \\ &= \int_0^\infty \exp(-x) \sum_0^\infty \frac{a_k}{k!} (x\lambda)^k dx \\ &= \int_0^\infty F(\lambda x) \exp(-x) dx. \end{aligned}$$

A practical problem with perturbation theory is that, apart from a few classroom examples, one is able to calculate only the lower order terms of the perturbation series. Although in principle it is impossible to divine the rest of a series by knowing its terms through a given order, a technique which in some cases turned out to work is that of Padé approximants [1,32]. A Padé  $[M, N]$  approximant to a series is a rational function

$$R_{MN}(z) = \frac{P_M(z)}{Q_N(z)} \quad (104)$$

whose power expansion near  $z = 0$  is equal to the first  $M + N$  terms of the series. It has been proved [31] that the Padé  $[N, N]$  approximants converge to the true eigenvalue of the anharmonic oscillator with  $x^4$  or  $x^6$  perturbation. The Padé  $[M, N]$  approximant to a function  $f(z)$

is unique, but its domain of analyticity is generally larger. Even for asymptotic series whose first terms are known one can write the Padé approximants. One can either use directly the Padé approximant as the value of  $E(\lambda)$  for the desired value of  $\lambda$ , or can insert it into the Borel summation method. For the case of the quartic anharmonic oscillator (Eq. (100)) both methods have been proved to work (at the cost of calculating some tens of terms of the series).

Another approach to the problem is the method of self-similar approximants [55], whereby approximants to the function  $E(\lambda)$  for which the asymptotic series is known are sought by means of products

$$f_{2p}(\lambda) = \prod_{i=1}^p (1 + A_i \lambda)^{n_i}. \quad (105)$$

The  $2p$  parameters  $A_i, n_i, 1 \leq i \leq p$ , are determined by equating the Taylor expansion of  $f_{2p}(\lambda)$  with the asymptotic series through order  $2p$  ( $a_0 = 1$  can be assumed, with no loss of generality, see [55]). Also, odd order approximants  $f_{2p+1}$  are possible. For the anharmonic oscillator (Eq. (100)) the calculations exhibit a steady convergence to the correct value of the energy of both the even order and the odd order approximants also for  $\lambda = 200$ .

The problem with the above approaches is that their efficiency seems limited to toy models as the anharmonic oscillator. For realistic problems it is difficult to establish in advance that the method converges to the correct answer.

## Perturbation of the Continuous Spectrum

In this section we consider the effect of a perturbing potential  $V$  on states belonging to the continuous spectrum. Since the problem is interesting mainly for the theory of scattering, we will assume that the unperturbed Hamiltonian  $H_0$  is the free Hamiltonian of a particle of mass  $m$ . Also, assuming that the potential  $V(\vec{r})$  vanishes at infinity, the spectrum of the free Hamiltonian  $H_0$  and the continuous spectrum of the exact Hamiltonian  $H = H_0 + \lambda V$  are equal and consist of the positive real semi-axis. Given an energy  $E = \hbar^2 k^2 / 2m$ , the problem is how the potential  $V$  affects that particular eigenfunction  $\psi_0$  of  $H_0$  which would represent the state of the system if the interaction potential were absent.

Letting  $\psi = \psi_0 + \delta\psi$ , the Schrödinger equation reads

$$(E - H_0)\delta\psi = \lambda V(\psi_0 + \delta\psi). \quad (106)$$

In the spirit of the perturbation approach,  $\delta\psi$  can be calculated by an iterative process provided we are able to find the solution of the inhomogeneous equation

$$(E - H_0)\delta\psi = \zeta \quad (107)$$

in the form

$$\delta\psi = \widetilde{G}_0 \zeta, \quad (108)$$

$\widetilde{G}_0$  being the Green's function of Eq. (107). Assuming that  $\widetilde{G}_0$  is known, we find

$$\begin{aligned} \delta\psi &= \lambda \widetilde{G}_0 V \psi \\ &= \lambda \widetilde{G}_0 V \psi_0 + \lambda \widetilde{G}_0 V \delta\psi \\ &= \lambda \widetilde{G}_0 V \psi_0 + \lambda \widetilde{G}_0 V (\lambda \widetilde{G}_0 V \psi_0 + \lambda \widetilde{G}_0 V \delta\psi) \\ &= \lambda \widetilde{G}_0 V \psi_0 + \lambda^2 \widetilde{G}_0 V \widetilde{G}_0 V \psi_0 + \lambda^2 \widetilde{G}_0 V \widetilde{G}_0 V \delta\psi \\ &= \dots, \end{aligned} \quad (109)$$

that is  $\delta\psi$  is written as a power expansion in  $\lambda$  in terms of the free wave function  $\psi_0$ .

### Scattering Solutions and Scattering Amplitude

One has to decide which eigenfunction of  $H_0$  must be inserted into the above expression, and which Green function  $\widetilde{G}_0$  must be used, since, of course, the solution of Eq. (107) is not unique. The questions are strongly interrelated, and the answers depend on which solution of the exact Schrödinger equation one wishes to find. Since the study of the perturbation of the continuous spectrum is relevant mainly for the theory of potential scattering, we will focus on this aspect. In the theory of scattering it is shown [24] that, for a potential  $V(\vec{r})$  vanishing faster than  $1/r$  for  $r \rightarrow \infty$ , a wave function  $\psi$  which in the asymptotic region is an eigenfunction of the momentum operator plus an outgoing wave

$$\psi \xrightarrow{r \rightarrow \infty} \exp(i\vec{k} \cdot \vec{r}) + f_{\vec{k}}(\theta, \varphi) \frac{\exp(ikr)}{r} \quad (110)$$

( $\theta, \varphi$  being the polar angles with respect to the  $\vec{k}$  axis) is suitable for describing the process of diffusion of a beam of free particles with momentum  $\vec{k}$  which impinge onto the interaction region and are scattered according the amplitude  $f_{\vec{k}}(\theta, \varphi)$ . (The character of outgoing wave of the second term in Eq. (110) is apparent when the time factor  $\exp(-iEt)$  is taken into account.) The differential cross section  $d\sigma/d\Omega$  is the ratio of the flux of the probability current density due to the outgoing wave to the flux due to the impinging plane wave. One finds

$$\frac{d\sigma}{d\Omega} = |f_{\vec{k}}(\theta, \varphi)|^2. \quad (111)$$

In conclusion, we require that  $\psi_0$  is a plane wave, and that the Green function  $\widetilde{G}_0$  has to be chosen in such a way as to yield an outgoing wave for large  $r$ .

Thus we need to solve the equation

$$\begin{aligned} (\vec{k}^2 + \Delta) \delta\psi &= \lambda \frac{2m}{\hbar^2} V \left( \exp(i\vec{k} \cdot \vec{r}) + \delta\psi \right) \\ &\equiv U(\vec{r}) \left( \exp(i\vec{k} \cdot \vec{r}) + \delta\psi \right) \end{aligned} \quad (112)$$

with the asymptotic condition  $\delta\psi \rightarrow f_{\vec{k}} \exp(ikr)/r$  for  $r \rightarrow \infty$ . In terms of the Green's function  $G_0(\vec{r}, \vec{r}')$ , which satisfies the equation

$$(\Delta + \vec{k}^2) G_0(\vec{r}, \vec{r}') = \delta(\vec{r} - \vec{r}'), \quad (113)$$

the solution of Eq. (112) can be written as

$$\begin{aligned} \delta\psi(\vec{r}) &= \int G_0(\vec{r}, \vec{r}') U(\vec{r}') \\ &\quad \left[ \exp(i\vec{k} \cdot \vec{r}') + \delta\psi(\vec{r}') \right] d\vec{r}', \end{aligned} \quad (114)$$

which is a form of the Lippmann–Schwinger equation [30]. The integral operator  $G_0$  with kernel  $G_0(\vec{r}, \vec{r}')$  is connected to the operator  $\widetilde{G}_0$  of Eq. (108) by the equation  $\widetilde{G}_0 = 2mG_0/\hbar^2$ .

The leading term of  $\delta\psi$  for  $r \rightarrow \infty$  is determined by the leading term of  $G_0(\vec{r}, \vec{r}')$ , so we look for a solution of Eq. (113) with the behavior of outgoing wave for  $r \rightarrow \infty$ . Due to translation and rotation invariance (if both the incoming beam and the scattering potential are translated or rotated by the same amount, the scattering amplitude  $f_{\vec{k}}(\theta, \varphi)$  is unchanged), we require for the solution a dependence only on  $|\vec{r} - \vec{r}'|$ .

Recalling that  $\Delta 1/r = -4\pi\delta(\vec{r})$ , we look for a solution of Eq. (113) with  $\vec{r}' = 0$  of the form  $-F(r)/(4\pi r)$ , with  $F(0) = 1$ . The function  $G_0(\vec{r}, \vec{r}')$  then will be

$$G_0(\vec{r}, \vec{r}') = \frac{F(|\vec{r} - \vec{r}'|)}{|\vec{r} - \vec{r}'|}. \quad (115)$$

The equation for  $F(r)$  is

$$F'' + k^2 F = 0, \quad (116)$$

whose solutions are  $\exp(\pm ikr)$  (outgoing and incoming wave respectively). In conclusion for  $G_0$  we find

$$G_0(\vec{r}, \vec{r}') = -\frac{1}{4\pi} \frac{\exp(ik|\vec{r} - \vec{r}'|)}{|\vec{r} - \vec{r}'|}. \quad (117)$$

The solution of the Schrödinger equation with the Green function given in Eq. (117) is denoted as  $\psi_{\vec{k}}^+$  and obeys the integral equation known as the Lippmann–Schwinger



equation [30]:

$$\psi_{\vec{k}}^+(\vec{r}) = \exp(i\vec{k} \cdot \vec{r}) - \frac{1}{4\pi} \int \frac{\exp(i\vec{k} |\vec{r} - \vec{r}'|)}{|\vec{r} - \vec{r}'|} U(\vec{r}') \psi_{\vec{k}}^+(\vec{r}') d\vec{r}'. \quad (118)$$

The behavior for  $r \rightarrow \infty$  can be easily checked to be as in Eq. (110) by inserting the expansion

$$|\vec{r} - \vec{r}'| = r - \frac{\vec{r} \cdot \vec{r}'}{r} + O(1/r) \quad (119)$$

into the Green function  $G_0$ . We find ( $\hat{r} \equiv \vec{r}/r$ )

$$-\frac{1}{4\pi} \frac{\exp(i\vec{k} |\vec{r} - \vec{r}'|)}{|\vec{r} - \vec{r}'|} \xrightarrow{r \rightarrow \infty} -\frac{1}{4\pi} \frac{\exp[i\vec{k} (r - \hat{r} \cdot \vec{r}')] }{r} \left[ 1 + \frac{\vec{r} \cdot \vec{r}'}{r^2} \right], \quad (120)$$

which yields for  $\psi_{\vec{k}}^+(\vec{r})$  ( $\vec{k}_f \equiv k\hat{r}$ )

$$\psi_{\vec{k}}^+(\vec{r}) \xrightarrow{r \rightarrow \infty} \exp(i\vec{k} \cdot \vec{r}) - \frac{1}{4\pi} \frac{\exp(ikr)}{r} \int \exp(-i\vec{k}_f \cdot \vec{r}') U(\vec{r}') \psi_{\vec{k}}^+(\vec{r}') d\vec{r}'. \quad (121)$$

The solutions  $\psi_{\vec{k}}^+(\vec{r})$  are normalized as the plane waves  $\exp(i\vec{k} \cdot \vec{r})$ :

$$(\psi_{\vec{k}}^+, \psi_{\vec{k}'}^+) = (2\pi)^3 \delta(\vec{k} - \vec{k}'). \quad (122)$$

In addition, they are orthogonal to any possible bound state solution of the Schrödinger equation with the Hamiltonian  $H = H_0 + \lambda V$ . Together with the bound state solutions they constitute a complete set. On a par with the solutions  $\psi_{\vec{k}}^+(\vec{r})$  one can also envisage solutions  $\psi_{\vec{k}}^-(\vec{r})$  with asymptotic behavior of incoming wave. They are obtained using for  $H$  (see Eq. (116)) the solution  $\exp(-ikr)$ . The normalization and orthogonality properties of the functions  $\psi_{\vec{k}}^-(\vec{r})$  are the same as for the  $\psi_{\vec{k}}^+(\vec{r})$  functions.

From Eq. (121) we derive an implicit expression for the scattering amplitude  $f_{\vec{k}}(\theta, \varphi)$ :

$$f_{\vec{k}}(\theta, \varphi) = -\frac{1}{4\pi} \int \exp(-i\vec{k}_f \cdot \vec{r}') U(\vec{r}') \psi_{\vec{k}}^+(\vec{r}') d\vec{r}' \quad (123)$$

where the unknown function  $\psi_{\vec{k}}^+(\vec{r})$  still appears. Letting  $\varphi_f \equiv \exp(i\vec{k}_f \cdot \vec{r})$ , Eq. (123) can also be written as

$$f_{\vec{k}}(\theta, \varphi) = -\frac{1}{4\pi} \left( \varphi_f, U \psi_{\vec{k}}^+ \right). \quad (124)$$

## The Born Series and its Convergence

Equations (118) and (124) are the starting point to obtain the expression of the exact wave function  $\psi_{\vec{k}}^+(\vec{r})$  and the scattering amplitude  $f_{\vec{k}}(\theta, \varphi)$  as a power series in  $\lambda$ , in the spirit of the perturbation approach. Recalling that  $U = (2m/\hbar^2)V$  (see Eq. (112)), if  $\varphi_{\vec{k}} \equiv \exp(i\vec{k} \cdot \vec{r})$  for  $\psi_{\vec{k}}^+(\vec{r})$  we find

$$\begin{aligned} \psi_{\vec{k}}^+ &= \varphi_{\vec{k}} + \lambda G_0 \frac{2m}{\hbar^2} V \varphi_{\vec{k}} + \lambda^2 G_0 \frac{2m}{\hbar^2} V G_0 \frac{2m}{\hbar^2} V \varphi_{\vec{k}} + \dots \\ &= \exp(i\vec{k} \cdot \vec{r}) + \lambda \int G_0(\vec{r}, \vec{r}') \frac{2m}{\hbar^2} V(\vec{r}') \\ &\quad \exp(i\vec{k} \cdot \vec{r}') d\vec{r}' + \lambda^2 \int d\vec{r}' \int d\vec{r}'' G_0(\vec{r}, \vec{r}') \frac{2m}{\hbar^2} \\ &\quad V(\vec{r}') G_0(\vec{r}', \vec{r}'') \frac{2m}{\hbar^2} V(\vec{r}'') \exp(i\vec{k} \cdot \vec{r}'') + \dots \end{aligned} \quad (125)$$

Inserting the above expansion into Eq. (124), for the scattering amplitude  $f_{\vec{k}}(\theta, \varphi)$  we find

$$f_{\vec{k}}(\theta, \varphi) = \sum_{n=1}^{\infty} f_{\vec{k}}^{(n)}(\theta, \varphi) \quad (126)$$

where  $f_{\vec{k}}^{(n)}$ , the contribution of order  $n$  in  $\lambda$  to  $f_{\vec{k}}(\theta, \varphi)$ , is obtained by substituting  $\psi_{\vec{k}}^+$  in Eq. (124) with the contribution of order  $n-1$  of the expansion Eq. (125). The term of order 1 is called the Born approximation [8], and is given by

$$\begin{aligned} f_{\vec{k}}^B(\theta, \varphi) &\equiv f_{\vec{k}}^{(1)}(\theta, \varphi) = -\frac{1}{4\pi} \left( \frac{2m\lambda}{\hbar^2} \right) \\ &\quad \int \exp[i(\vec{k} - \vec{k}_f) \cdot \vec{r}'] V(\vec{r}') d\vec{r}'. \end{aligned} \quad (127)$$

The term of order 2 is

$$f_{\vec{k}}^{(2)}(\theta, \varphi) = -\frac{1}{4\pi} \left( \frac{2m\lambda}{\hbar^2} \right)^2 (\varphi_f, V G_0 V \varphi_{\vec{k}}) \quad (128)$$

and the general term of order  $n$  is

$$\begin{aligned} f_{\vec{k}}^{(n)}(\theta, \varphi) &= -\frac{1}{4\pi} \left( \frac{2m\lambda}{\hbar^2} \right)^n (\varphi_f, V G_0 V \dots G_0 V \varphi_{\vec{k}}) \\ &\quad (n \text{ times } V). \end{aligned} \quad (129)$$

The scattering amplitude through order  $n$  is

$$f_{\vec{k}}^{[n]}(\theta, \varphi) = \sum_{i=1}^n f_{\vec{k}}^{(i)}(\theta, \varphi) \quad (130)$$

with  $f_k^{(1)}(\theta, \varphi) \equiv f_k^B(\theta, \varphi)$ , and the series  $\sum_{i=1}^{\infty} f_k^{(i)}(\theta, \varphi)$  is known as the Born series [8]. Of course, when using Eq. (130) for calculating the differential cross section  $d\sigma/d\Omega$  only terms of order not exceeding  $n$  should be consistently retained.

For a discussion of the range of validity and the convergence of the expansions Eqs. (125) and (126) it is convenient to pose the problem in the framework of integral equations in the Hilbert space  $L^2$  [40,54], which provides a natural notion of convergence. To this purpose, since  $\psi_k^+(\vec{r})$  is not square integrable, we start assuming that the potential  $V(\vec{r})$  is summable

$$\int |V(\vec{r})| d\vec{r} < \infty \quad (131)$$

and multiply Eq. (118) by  $|V(\vec{r})|^{\frac{1}{2}}$  [20,41,45]. Letting  $\epsilon_V(\vec{r}) \equiv V(\vec{r})/|V(\vec{r})|$  ( $\epsilon_V(\vec{r}) \equiv 0$  if  $V(\vec{r}) = 0$ ) and defining

$$\zeta_k^+(\vec{r}) \equiv |V(\vec{r})|^{\frac{1}{2}} \psi_k^+(\vec{r}) \quad (132)$$

$$K'(\vec{r}, \vec{r}') \equiv -\frac{2m}{\hbar^2} G_0(\vec{r}, \vec{r}') |V(\vec{r})|^{\frac{1}{2}} |V(\vec{r}')|^{\frac{1}{2}} \epsilon_V(\vec{r}'), \quad (133)$$

Eq. (118) reads:

$$\begin{aligned} \zeta_k^+(\vec{r}) &= |V(\vec{r})|^{\frac{1}{2}} \exp(i\vec{k} \cdot \vec{r}) \\ &+ \lambda \int K'(\vec{r}, \vec{r}') \zeta_k^+(\vec{r}') d\vec{r}'. \end{aligned} \quad (134)$$

Now the function in front of the integral is square integrable and the kernel  $K'(\vec{r}, \vec{r}')$  is square integrable too

$$\begin{aligned} \int d\vec{r} \int d\vec{r}' |K'(\vec{r}, \vec{r}')|^2 \\ = \int d\vec{r} \int d\vec{r}' \frac{|V(\vec{r})| |V(\vec{r}')|}{|\vec{r} - \vec{r}'|^2} < \infty \end{aligned} \quad (135)$$

provided the potential  $V(\vec{r})$  obeys the additional condition

$$\int d\vec{r}' \frac{|V(\vec{r}')|}{|\vec{r} - \vec{r}'|^2} < \infty. \quad (136)$$

Equation (134) can be formally written as

$$\zeta_k^+ = \zeta_0 + \lambda \hat{K}' \zeta_k^+, \quad \zeta_0 \equiv |V(\vec{r})|^{\frac{1}{2}} \exp(i\vec{k} \cdot \vec{r}), \quad (137)$$

$\hat{K}'$  being the integral operator with kernel  $K'$  given in Eq. (133). The function  $\zeta_k^+$  is formally given as

$$\zeta_k^+ = (I - \lambda \hat{K}')^{-1} \zeta_0 \quad (138)$$

where the inverse operator  $(I - \lambda \hat{K}')^{-1}$  exists except for those values of  $\lambda$  (singular values) for which  $I - \lambda \hat{K}'$  has the eigenvalue 0.

Since by Eq. (135)  $\hat{K}'$  is a compact operator, the singular values are isolated points which obey the inequality  $|\lambda| \geq \|\hat{K}'\|^{-1}$ , since the spectrum of an operator is contained in the closed disc of radius equal to the norm of the operator. Thus, when  $\|\lambda \hat{K}'\| < 1$  the inverse operator  $(I - \lambda \hat{K}')^{-1}$  exists and is given by the Neumann series

$$(I - \lambda \hat{K}')^{-1} = I + \lambda \hat{K}' + \lambda^2 \hat{K}'^2 + \dots \equiv I + R_{\lambda}^{K'}, \quad (139)$$

which is clearly norm convergent. By the inequality

$$\begin{aligned} \|\lambda \hat{K}'\|^2 &\leq \lambda^2 \text{Tr}(\hat{K}'^\dagger \hat{K}') \\ &= \lambda^2 \int d\vec{r} \int d\vec{r}' |K'(\vec{r}, \vec{r}')|^2 \end{aligned} \quad (140)$$

we see that, if

$$\lambda^2 \frac{m^2}{4\pi^2 \hbar^4} \int d\vec{r} \int d\vec{r}' \frac{|V(\vec{r})| |V(\vec{r}')|}{|\vec{r} - \vec{r}'|^2} < 1, \quad (141)$$

the condition  $\|\lambda \hat{K}'\| < 1$  is satisfied and consequently the inverse operator  $(I - \lambda \hat{K}')^{-1}$  exists. In conclusion, a sufficient condition for the convergence of the expansion

$$\zeta_k^+ = \zeta_0 + \lambda \hat{K}' \zeta_0 + \lambda^2 \hat{K}'^2 \zeta_0 + \dots \quad (142)$$

in the  $L^2$  norm is that Eq. (141) holds [43]. Since  $\|\hat{K}'\|^4 \leq \text{Tr}(\hat{K}'^\dagger \hat{K}' \hat{K}'^\dagger \hat{K}')$ , by the Riemann–Lebesgue lemma it is possible to prove [56] that for any given  $\lambda$  the condition  $\|\lambda \hat{K}'\| < 1$  is satisfied provided the energy  $\hbar^2 k^2/2m$  is sufficiently large.

The implications for the convergence of the expansion Eq. (126) of the scattering amplitude are immediate, once Eq. (124) is written in the form

$$\begin{aligned} f_k^-(\theta, \varphi) &= -\lambda \frac{m}{2\pi \hbar^2} \\ &\left( |V(\vec{r})|^{\frac{1}{2}} \varphi_f(\vec{r}) \epsilon_V(\vec{r}), |V(\vec{r})|^{\frac{1}{2}} \psi_k^+(\vec{r}) \right). \end{aligned} \quad (143)$$

The Born series converges whenever the iterative solution of Eq. (137) converges, that is if Eq. (135) is satisfied. As noted above, for any given  $\lambda$  this happens for sufficiently large energy. An additional useful result is that the Born approximation  $f_k^B(\theta, \varphi)$  (or the expansion  $f_k^{[n]}(\theta, \varphi)$  through any  $n$ ) converges to the exact scattering amplitude  $f_k^-(\theta, \varphi)$  when the energy grows to infinity. More precisely [48], if Eq. (131) holds then

$$\left| f_k^-(\theta, \varphi) - f_k^{[n]}(\theta, \varphi) \right| \xrightarrow{k \rightarrow \infty} 0. \quad (144)$$

If  $\|\lambda \hat{K}'\| \geq 1$  the Neumann series Eq. (139) does not converge and the perturbation approach is no longer viable. However, if  $\lambda$  is not a singular value Eq. (137) can be solved by reducing it to an integral equation with a kernel  $D$  of norm less than 1 plus a problem of linear algebra [54]. In fact, for any positive value  $L$  the operator  $\hat{K}'$  can be approximated by a finite rank operator  $F$

$$F\zeta = \sum_{i=1}^n \alpha_i(\vec{r}) (\beta_i(\vec{r}'), \zeta(\vec{r}')) \quad (145)$$

such that, if  $D \equiv \hat{K}' - F$ ,  $\|D\| < 1/L$ . Equation (137) then reads

$$(I - \lambda D)\zeta_k^+ = \zeta_0 + \lambda F\zeta_k^+. \quad (146)$$

Since for  $|\lambda| < L$  we have  $\|\lambda D\| < 1$ ,  $\zeta_k^+$  can be written in terms of the appropriate Neumann series  $I + R_\lambda^D$  (see Eq. (139)):

$$\zeta_k^+ = \zeta_0 + R_\lambda^D \zeta_0 + F\zeta_k^+ + R_\lambda^D F\zeta_k^+. \quad (147)$$

The unknown quantities  $(\beta_i, \zeta_k^+)$  which appear in the RHS of Eq. (147) are determined by solving the linear-algebraic problem obtained by left-multiplying both sides of the equation by  $\beta_r$ ,  $1 \leq r \leq n$ . The values of  $\lambda$  in the range  $|\lambda| < L$  for which the algebraic problem is not soluble are the singular values of Eq. (137) in that range. Thus, Eq. (137) can be solved for any non-singular value.

### Time Dependent Perturbations

A rather different problem is presented by the case that a time independent Hamiltonian  $H_0$ , for which the spectrum and the eigenfunctions are known, is perturbed by a time dependent potential  $V(t)$ . This occurs, for example, when an atom or a molecule interacts with an external electromagnetic field. For the total Hamiltonian  $H = H_0 + \lambda V(t)$  stationary states no longer exist, and the relevant question is how the perturbation affects the time evolution of the system. We assume that the state  $\psi$  is known at a given time, which can be chosen as  $t = 0$ , and search for  $\psi(t)$ . Obviously, any time  $t_0$  prior to the setting on of the perturbation  $\lambda V(t)$  could be chosen instead of  $t = 0$ .

The time dependent Schrödinger equation reads

$$i\hbar \frac{\partial \psi}{\partial t} = H\psi(t) = H_0\psi(t) + \lambda V(t)\psi(t). \quad (148)$$

At any  $t$ ,  $\psi(t)$  can be expanded in the basis of the eigenfunctions  $\varphi_n(t)$  of  $H_0$ :

$$H_0\varphi_n(t) = E_n\varphi_n(t) \quad (149)$$

$$\varphi_n(t) = \varphi_n(0) \exp(-iE_n t/\hbar) \equiv \zeta_n \exp(-iE_n t/\hbar). \quad (150)$$

For the sake of simplicity we treat  $H_0$  as if it only had discrete spectrum, but the presence of a continuous component of the spectrum does not create any problem.

We can write [11,42]

$$\psi(t) = \sum_n a_n(t) \varphi_n(t). \quad (151)$$

Note that the basis vectors  $\varphi_n(t)$  are themselves time dependent (by the phase factor given in Eq. (150)), whereas the vectors  $\zeta_n$  are time independent. The isolation of the contribution of  $H_0$  to the time evolution as a time dependent factor allows a simpler system of equations for the unknown coefficients  $a_n(t)$ .

Substituting expansion Eq. (151) into Eq. (148) we find

$$\begin{aligned} i\hbar \sum_n \dot{a}_n(t) \varphi_n(t) + \sum_n a_n(t) E_n \varphi_n(t) \\ = \sum_n a_n(t) E_n \varphi_n(t) + \lambda \sum_n a_n(t) V(t) \varphi_n(t). \end{aligned}$$

By left multiplying by  $\varphi_k(t)$ , for the coefficients  $a_k(t)$  we find the system of equations

$$\begin{aligned} i\hbar \dot{a}_k(t) &= \lambda \sum_n a_n(t) (\varphi_k(t), V(t) \varphi_n(t)) \\ &\equiv \lambda \sum_n V_{kn}^I(t) a_n(t), \end{aligned} \quad (152)$$

where we have defined

$$V_{kn}^I(t) = (\varphi_k(t), V(t) \varphi_n(t)). \quad (153)$$

The matrix elements  $V_{kn}^I(t)$  are the matrix elements of an operator  $V^I(t)$  between the time independent vectors  $\zeta_k, \zeta_n$ . Indeed, since

$$\varphi_n(t) = \zeta_n \exp(-iE_n t/\hbar) = \exp(-iH_0 t/\hbar) \zeta_n, \quad (154)$$

Eq. (153) can be written as

$$\begin{aligned} V_{kn}^I(t) &= (\exp(-iH_0 t/\hbar) \zeta_k, V(t) \exp(-iH_0 t/\hbar) \zeta_n) \\ &\equiv (\zeta_k, V^I(t) \zeta_n), \end{aligned} \quad (155)$$

where the operator  $V^I(t)$  is defined as follows:

$$V^I(t) \equiv \exp(iH_0 t/\hbar) V(t) \exp(-iH_0 t/\hbar). \quad (156)$$

System Eq. (152) must be supplemented with the appropriate initial conditions, which depend on the particular problem. The commonest application of time dependent perturbation theory is the calculation of transition

probabilities between eigenstates of  $H_0$ . Thus, we assume that at  $t = 0$  the system is in an eigenstate of the unperturbed Hamiltonian  $H_0$ , say the state  $\varphi_1$ . In this case  $a_1(0) = 1$ ,  $a_n(0) = 0$  if  $n \neq 1$ .

In the spirit of perturbation theory, each  $a_n$  is expanded into powers of  $\lambda$

$$\begin{aligned} a_1(t) &= 1 + \sum_{r=1} \lambda^r a_1^{(r)}(t), \\ a_n(t) &= \sum_{r=1} \lambda^r a_n^{(r)}(t) \quad n \neq 1, \end{aligned} \quad (157)$$

and terms of equal order are equated. For  $a_k^{(r)}$  we find

$$i\hbar \dot{a}_k^{(r)} = \sum_n V_{kn}^I a_n^{(r-1)}, \quad r > 0. \quad (158)$$

By Eq. (157), for any  $r > 0$ ,  $a_n^{(r)}(0) = 0$ . As a consequence, for  $r = 1$  we have

$$\begin{aligned} a_k^{(1)} &= \frac{-i}{\hbar} \int_0^t V_{k1}^I(t_1) dt_1 \\ &= \frac{-i}{\hbar} \int_0^t (\zeta_k, V(t_1) \zeta_1) \exp(i\Delta E_{k1} t_1 / \hbar) dt_1, \end{aligned} \quad (159)$$

where  $\Delta E_{k1} \equiv E_k - E_1$ . For  $r = 2$  we find

$$\begin{aligned} i\hbar \dot{a}_k^{(2)} &= \sum_n V_{kn}^I(t) a_n^{(1)}(t) \\ &= \frac{-i}{\hbar} \sum_n V_{kn}^I(t) \int_0^t V_{n1}^I(t_1) dt_1 \end{aligned}$$

whose solution is

$$\begin{aligned} a_k^{(2)}(t) &= \left(\frac{-i}{\hbar}\right)^2 \int_0^t dt_2 \int_0^{t_2} dt_1 \sum_n V_{kn}^I(t_2) V_{n1}^I(t_1) \\ &= \left(\frac{-i}{\hbar}\right)^2 \int_0^t dt_2 \int_0^{t_2} dt_1 \sum_n (\zeta_k, V(t_2) \zeta_n) \\ &\quad \exp(i\Delta E_{kn} t_2 / \hbar) (\zeta_n, V(t_1) \zeta_1) \exp(i\Delta E_{n1} t_1 / \hbar). \end{aligned} \quad (160)$$

It is clear how the calculation proceeds for higher values of  $r$ . The general expression is

$$\begin{aligned} a_k^{(r)}(t) &= \left(\frac{-i}{\hbar}\right)^r \int_0^t dt_r \int_0^{t_r} dt_{r-1} \cdots \int_0^{t_3} dt_2 \\ &\quad \int_0^{t_2} dt_1 \sum V_{kn_r}^I(t_r) V_{n_r n_{r-1}}^I(t_{r-1}) \cdots \\ &\quad V_{n_3 n_2}^I(t_2) V_{n_2 1}^I(t_1). \end{aligned} \quad (161)$$

By the completeness of the vectors  $\zeta_n$ , the sums over the intermediate states can be substituted by the identity and the expression of  $a_k^{(r)}$  is simplified into

$$\begin{aligned} a_k^{(r)}(t) &= \left(\frac{-i}{\hbar}\right)^r \int_0^t dt_r \int_0^{t_r} dt_{r-1} \cdots \int_0^{t_3} dt_2 \\ &\quad \int_0^{t_2} dt_1 (\zeta_k, V^I(t_r) V^I(t_{r-1}) \cdots V^I(t_2) V^I(t_1) \zeta_1). \end{aligned} \quad (162)$$

It is customary to write Eq. (162) in a different way. The  $r$ -dimensional cube  $0 \leq t_i \leq t$ ,  $1 \leq i \leq r$ , can be split into  $r!$  subdomains

$$0 \leq t_{p_1} \leq t_{p_2} \leq \cdots \leq t_{p_{r-1}} \leq t_{p_r} \leq t, \quad (163)$$

with  $\{p_1, p_2, \dots, p_{r-1}, p_r\}$  a permutation of  $\{1, 2, \dots, r-1, r\}$ . The time ordered product of  $r$  (non-commuting) operators  $V^I(t_{p_1}), V^I(t_{p_2}), \dots, V^I(t_{p_r})$  is introduced according to the definition

$$\begin{aligned} T[V^I(t_{p_1}) \cdots V^I(t_{p_r})] &\equiv V^I(t_r) \cdots V^I(t_1), \\ t_1 \leq t_2 \leq \cdots \leq t_r. \end{aligned} \quad (164)$$

If  $(-i\lambda/\hbar)^r (\zeta_k, T[V^I(t_{p_1}) \cdots V^I(t_{p_r})] \zeta_1)$  is integrated over the  $r$ -cube, then each of the  $r!$  subdomains defined by Eq. (163) yields the same contribution. As a consequence Eq. (161) can be written as

$$\begin{aligned} a_k^{(r)}(t) &= \frac{1}{r!} \left(\frac{-i}{\hbar}\right)^r \int_0^t dt_r \int_0^{t_r} dt_{r-1} \cdots \int_0^{t_3} dt_2 \\ &\quad \int_0^{t_2} dt_1 (\zeta_k, T[V^I(t_r) V^I(t_{r-1}) \cdots V^I(t_2) V^I(t_1)] \zeta_1). \end{aligned} \quad (165)$$

The amplitudes  $a_k(t)$  can then be written as

$$a_k(t) = \left(\zeta_k, T\left[\exp\left(\frac{-i\lambda}{\hbar}\right) \int_0^t V^I(t') dt'\right] \zeta_1\right) \quad (166)$$

with obvious significance of the  $T$ -exponential: each monomial in the  $V^I$  operators which appear in the expansion of the exponential is to be time ordered according to the  $T$ -prescription. If the initial state is given at time  $t_0$  the integral appearing in the  $T$ -exponential should start at  $t_0$ . We define

$$U^I(t, t_0) \equiv T\left[\exp\left(\frac{-i\lambda}{\hbar}\right) \int_{t_0}^t V^I(t') dt'\right]. \quad (167)$$

The expansion of the  $T$ -exponential into monomials in the  $V^I$  operators is called the Dyson series [13,14]. It is extensively employed in perturbative quantum field theory.

From Eqs. (166) and (167) it is easy to derive an expression for the time evolution operator  $U(t, t_0)$  such that

$$U(t, t_0)\psi(t_0) = \psi(t). \quad (168)$$

Indeed,  $\psi(t)$  and  $\psi(t_0)$  can be expanded in the basis of the vectors  $\varphi_n(t)$  and  $\varphi_n(t_0)$  respectively as in Eq. (151). By the linearity of the Schrödinger equation it suffices to determine  $(\varphi_k(t), U(t, t_0)\varphi_n(t_0))$ , which we already know to be  $(\zeta_k, U^I(t, t_0)\zeta_n)$ . We have

$$\begin{aligned} &(\varphi_k(t), U(t, t_0)\varphi_n(t_0)) \\ &= (\exp(-iH_0 t/\hbar)\zeta_k, U(t, t_0)\exp(-iH_0 t_0/\hbar)\zeta_n) \\ &= (\zeta_k, \exp(iH_0 t/\hbar)U(t, t_0)\exp(-iH_0 t_0/\hbar)\zeta_n). \end{aligned}$$

As a consequence we find the equation

$$\begin{aligned} U(t, t_0) \\ = \exp(-iH_0 t/\hbar) T \left[ \exp\left(\frac{-i\lambda}{\hbar}\right) \int_{t_0}^t V^I(t') dt' \right] \\ \exp(iH_0 t_0/\hbar), \quad (169) \end{aligned}$$

which provides the perturbation expansion of the evolution operator  $U(t, t_0)$  in powers of  $\lambda$ .

It can be proved that if the operator function  $V(t)$  is strongly continuous and the operators  $V(t)$  are bounded, then the expansion which defines the  $T$ -exponential is norm convergent to a unitary operator, as expected [35]. The restriction to bounded operators  $V(t)$  does not detract from the range of applications of Eqs. (166) and (169), since time dependent perturbation theory is almost exclusively used for treating interactions of a system with external fields, which generate bounded interactions.

### Future Directions

The long and honorable service of perturbation theory in every sector of quantum mechanics must be properly acknowledged. Its future is perhaps already in our past: the main achievement is its application to quantum field theory where, just to quote an example, the agreement between the measured value and the theoretical prediction of the electron magnetic moment anomaly to ten significant digits has no rivals.

Despite its successes, still perturbation theory is confronted with fundamental questions. In most of realistic problems it is unknown whether the perturbation series is convergent or at least asymptotic. In non-relativistic quantum mechanics this does not represent a practical problem since only a limited number of terms can be calculated, but in quantum field theory, where higher order terms are in

principle calculable, this calls for dedicated investigations. There, in particular, conditions for recovering the exact amplitudes from the first terms of the series by such techniques as the Padé approximants or the self similar approximants, and the estimate on the bound of the error, deserve further investigation.

Somewhat paradoxically, it can be said that the future of perturbation theory is in the non-perturbative results (analyticity domains, large coupling constant behavior, tunneling effect ...) – an issue where much work has already been done – since they have proved to be complementary to the use of perturbation theory.

### Bibliography

#### Primary Literature

1. Baker GA, Graves-Morris P (1996) Padé approximants. Cambridge Univ. Press, Cambridge
2. Bargmann V (1964) Note on Wigner's theorem on symmetry operations. J Math Phys 5:862–868
3. Bender CM, Wu TT (1969) Anharmonic oscillator. Phys Rev 184:1231–1260
4. Bloch C (1958) Sur la théorie des perturbations des états liés. Nucl Phys 6:329–347
5. Böhm A (1993) Quantum mechanics, foundations and applications. Springer, New York, pp 208–215
6. Borel E (1899) Mémoires sur le séries divergentes. Ann Sci École Norm Sup 16:9–136
7. Born M, Heisenberg W, Jordan P (1926) Zur Quantenmechanik, II. Z Phys 35:557–615
8. Born M (1926) Quantenmechanik der Stossvorgänge. Z Phys 38:803–827
9. Brillouin L (1932) Perturbation problem and self consistent field. J Phys Radium 3:373–389
10. Courant R, Hilbert D (1989) Methods of mathematical physics, vol I. Wiley, New York, pp 343–350
11. Dirac PAM (1926) On the theory of quantum mechanics. Proc Roy Soc A112:661–677
12. Dyson FJ (1952) Divergence of perturbation theory in quantum electrodynamics. Phys Rev 85:631–632
13. Dyson FJ (1949) The radiation theories of Tomonaga, Schwinger and Feynman. Phys Rev 75:486–502
14. Dyson FJ (1949) The S-matrix in quantum electrodynamics. Phys Rev 75:1736–1755
15. Epstein ST (1954) Note on perturbation theory. Amer J Phys 22:613–614
16. Epstein ST (1968) Uniqueness of the energy in perturbation theory. Amer J Phys 36:165–166
17. Feynman RP (1939) Forces in molecules. Phys Rev 56:340–343
18. Fock VA (1935) Zur Theorie des Wasserstoffatoms. Z Phys 98:145–154
19. Graffi S, Grecchi V, Simon B (1970) Borel summability: Application to the anharmonic oscillator. Phys Lett B32:631–634
20. Grossman A (1961) Schrödinger scattering amplitude I. J Math Phys 3:710–713
21. Hamermesh M (1989) Group theory and its application to physical problems. Dover, New York, pp 32–114



22. Hannabuss K (1997) Introduction to quantum theory. Clarendon, Oxford, pp 131–136
23. Hellmann H (1937) Einführung in die Quantenchemie. Deuticke, Leipzig
24. Joachain J (1983) Quantum collision theory. North-Holland, Amsterdam
25. Kato T (1966) Perturbation theory for linear operators. Springer, New York
26. Kato T (1949) On the convergence of the perturbation method I. *Progr Theor Phys* 4:514–523
27. Kramers HA (1957) Quantum mechanics. North-Holland, Amsterdam, pp 198–202
28. Krieger JB (1968) Asymptotic properties of perturbation theory. *J Math Phys* 9:432–435
29. Landau LD, Lifshitz EM (1960) Mechanics. Pergamon Press, Oxford, p 129
30. Lippmann BA, Schwinger J (1950) Variational principles for scattering processes I. *Phys Rev* 79:469–480
31. Loeffel J, Martin A, Wightman A, Simon B (1969) Padé approximants and the anharmonic oscillator. *Phys Lett B* 30:656–658
32. Padé H (1899) Sur la représentation approchée d'une fonction pour des fonctions rationnelles. *Ann Sci Éco Norm Sup Suppl* 9(3):1–93
33. Rayleigh JW (1894–1896) The theory of sound, vol I. Macmillan, London, pp 115–118
35. Reed M, Simon B (1975) Methods of modern mathematical physics, vol II. Academic Press, New York, pp 282–283
34. Reed M, Simon B (1978) Methods of modern mathematical physics, vol IV. Academic Press, New York, pp 10–44
36. Rellich F (1937) Störungstheorie der Spektralzerlegung I. *Math Ann* 113:600–619
37. Rellich F (1937) Störungstheorie der Spektralzerlegung II. *Math Ann* 113:677–685
38. Rellich F (1939) Störungstheorie der Spektralzerlegung III. *Math Ann* 116:555–570
39. Rellich F (1940) Störungstheorie der Spektralzerlegung IV. *Math Ann* 117:356–382
40. Riesz F, Sz-Nagy B (1968) Leçons d'analyse fonctionnelle. Gauthier-Villars, Paris, pp 143–188
41. Rollnik H (1956) Streumaxima und gebundene Zustände. *Z Phys* 145:639–653
42. Sakurai JJ (1967) Advanced quantum mechanics. Addison-Wesley, Reading, pp 39–40
43. Scadron M, Weinberg S, Wright J (1964) Functional analysis and scattering theory. *Phys Rev* 135:B202–B207
44. Schrödinger E (1926) Quantisierung als Eigenwertproblem. *Ann Phys* 80:437–490
45. Schwartz J (1960) Some non-self-adjoint operators. *Comm Pure Appl Math* 13:609–639
46. Simon B (1970) Coupling constant analyticity for the anharmonic oscillator. *Ann Phys* 58:76–136
47. Simon B (1991) Fifty years of eigenvalue perturbation theory. *Bull Am Math Soc* 24:303–319
48. Thirring W (2002) Quantum mathematical physics. Springer, Berlin, p 177
49. von Neumann J, Wigner E (1929) Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. *Phys Z* 30:467–470
50. Watson G (1912) A theory of asymptotic series. *Philos Trans Roy Soc Lon Ser A* 211:279–313
51. Weyl H (1931) The theory of groups and quantum mechanics. Dover, New York
52. Wigner EP (1935) On a modification of the Rayleigh–Schrödinger perturbation theory. *Math Natur Anz (Budapest)* 53:477–482
53. Wigner EP (1959) Group theory and its application to the quantum mechanics of atomic spectra. Academic Press, New York
54. Yosida K (1991) Lectures on differential and integral equations. Dover, New York, pp 115–131
55. Yukalov VI, Yukalova EP (2007) Methods of self similar factor approximants. *Phys Lett A* 368:341–347
56. Zemach C, Klein A (1958) The Born expansion in non-relativistic quantum theory I. *Nuovo Cimento* 10:1078–1087

## Books and Reviews

- Hirschfelder JO, Byers Brown W, Epstein ST (1964) Recent developments in perturbation theory. In: *Advances in quantum chemistry*, vol 1. Academic Press, New York, pp 255–374
- Killingbeck J (1977) Quantum-mechanical perturbation theory. *Rep Progr Phys* 40:963–1031
- Mayer I (2003) Simple theorems, proofs and derivations in quantum chemistry. Kluwer Academic/Plenum Publishers, New York, pp 69–120
- Morse PM, Feshbach H (1953) Methods of theoretical physics, part 2. McGraw-Hill, New York, pp 1001–1106
- Wilcox CH (1966) Perturbation theory and its applications in quantum mechanics. Wiley, New York

## Perturbation Theory, Semiclassical

ANDREA SACCHETTI

Dipartimento di Matematica Pura ed Applicata,  
Università di Modena e Reggio Emilia, Modena, Italy

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 The WKB Approximation  
 Semiclassical Approximation in Any Dimension  
 Propagation of Quantum Observables  
 Future Directions  
 Bibliography

## Glossary

**Agmon metric** In the classically forbidden region where the potential energy  $V(x)$  is larger than the total energy  $E$ , i.e.  $V(x) > E$ , we introduce a notion of distance based on the *Agmon metric* defined as  $[V(x) - E]dx^2$ , where  $dx^2$  is the usual Riemann metric. We emphasize that such an Agmon metric is the “semiclassical” equivalent of the “classical” Jacobi met-

ric  $[E - V(x)]dx^2$  introduced in classical mechanics in the classically permitted region where  $V(x) < E$ .

**Asymptotic series** The notion of *asymptotic series* goes back to Poincaré. We say that a formal power series  $\sum_r a_r z^{-r}$  is taken to be the *asymptotic power series* for a function  $f(z)$ , as  $|z| \rightarrow \infty$  in a given infinite sector  $S = \{z \in \mathbb{C} : \alpha \leq \arg z \leq \beta\}$ , if for any fixed  $N$  the remainder term

$$R_N(z) = f(z) - \sum_{r=0}^{N-1} a_r z^{-r}$$

is such that  $R_N(z) = \mathcal{O}(z^{-N})$ , that is

$$|R_N(z)| \leq C_N |z^{-N}|, \quad \forall z \in S,$$

for some positive constant  $C_N$  depending on  $N$ .

### Classically allowed and forbidden regions – turning

**points** We distinguish two different regions: the region  $V(x) < E$  where the *classical motion is allowed* and the region  $V(x) > E$  where the *classical motion is forbidden*. The points that separate these two regions, that is such that  $V(x) = E$ , will play a special role and they are named *turning points*.

**Semiclassical limit** Cornerstone of Quantum Mechanics is the time-dependent Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \Delta \psi + V(x)\psi, \quad (1)$$

where  $V(x)$  is assumed to be a real-valued function, usually it represents the potential energy, and  $m$  is the mass of the particle. The solution of this equation defines the density of probability  $|\psi(x, t)|^2$  to find the particle in some region of space. The parameter  $\hbar$  in Eq. (1) is related to Planck's constant  $h$

$$\begin{aligned} \hbar &= \frac{h}{2\pi} = 1.0545 \times 10^{-27} \text{ erg sec} \\ &= 6.5819 \times 10^{-22} \text{ MeV sec.} \end{aligned}$$

According to the correspondence principle, when Planck's constant can be considered small with respect to the other parameters, such as masses and distances, then quantum theory approaches classical Newton theory. Thus, roughly speaking, we expect that classical mechanics is contained in quantum mechanics as a limiting form (i. e.  $\hbar \rightarrow 0$ ). The limit of small  $\hbar$ , when compared with the other parameters, is the so-called *semiclassical limit*. We should emphasize that making Planck's constant small in Eq. (1) is a rather singular limit and many difficult mathematical problems occur.

**Stokes lines** There is some misunderstanding in the literature concerning the name of the curves  $\Im[\xi(z)] = 0$ , where  $\Im(\xi)$  denotes the imaginary part of

$$\xi(z) = \frac{i}{\hbar} \int_{x_E}^z \sqrt{E - V(q)} dq, \quad z \in \mathbb{C},$$

and  $x_E \in \mathbb{R}$  is a turning point (i. e.  $V(x_E) = E$ ), the potential  $V$  is assumed to be an analytic function in the complex plane. As usually physicists do, and in agreement with Stokes' original treatment, we adopt here the convention to name as *Stokes line* any path coming from the turning point  $x_E$  such that  $\Im[\xi(z)] = 0$ ; reserving the name of *anti-Stokes lines*, or *regular lines*, for the curves such that  $\Re[\xi(z)] = 0$ , where  $\Re(\xi)$  denotes the real part of  $\xi$  (we should emphasize that most mathematicians adopt the opposite rule, that is they call Stokes lines the paths such that  $\Re[\xi(z)] = 0$ ).

**Tunnel effect** At a real (simple) turning point  $x_E$  where  $E = V(x)$  (and  $V'(x) \neq 0$ ), classical particles incident from an accessible region [where  $E > V(x)$ ] reverse their velocity and return back; the adjacent region [where  $E < V(x)$ ] is forbidden to these particles according to classical mechanics. In fact, quantum mechanically exponentially growing or damped waves can exist in forbidden regions and a quantum particle can pass through a potential barrier. This is the so-called *tunnel effect*.

**WKB** The WKB method, named after the contributions independently given by Wentzel, Kramers and Brillouin, consists of connecting the approximate solutions of the time independent Schrödinger equation across turning points.

### Definition of the Subject

Several kinds of perturbation methods are commonly used in quantum mechanics ► [Perturbation Theory and Molecular Dynamics](#), ► [Perturbation Theory in Quantum Mechanics](#), ► [Perturbation Theory](#). This chapter deals with *semiclassical approximations* where expressions for energy levels and wave functions are obtained in the limiting case of small values of Planck's constant.

We emphasize that wave functions are highly singular as the parameter  $\hbar$  goes to zero. In fact the semiclassical limit is a singular perturbation problem; namely, Eq. (1) suffers a reduction of order setting  $\hbar$  equal to zero and the resulting equation is not a differential equation and does not give the classical limit correctly. Therefore, ordinary perturbation methods, which usually give energy lev-

els and wave-functions as convergent power series of the small parameter, cannot be applied.

The main goal of *semiclassical methods* consists of obtaining *asymptotic series*, in the limit of small  $\hbar$ , for the quantum-mechanical quantities. For instance, semiclassical methods permit us to solve the eigenvalues problem

$$-\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi = E\psi, \quad x \in \mathbb{R}^n, \quad n \geq 1, \quad (2)$$

where the energy levels  $E$  and the wave-functions  $\psi(x)$  are given by means of asymptotic series in the limit of vanishing  $\hbar$ , or to obtain a direct link between the time evolution of a quantum observable and the Hamiltonian flux of the associated classical quantity.

## Introduction

The first contributions to this theory, in the framework of Quantum Mechanics, go back to Wentzel (1926), Kramers (1926) and Brillouin (1926). In their papers they independently developed the determination of connection formulas linking exponential and oscillatory approximations of the solution of Eq. (2) in dimension  $n = 1$  across a turning point. This method, explained in Sect. [The WKB Approximation](#), is usually called WKB approximation, or also JWKB method to acknowledge that the approximate connection formula was previously discovered by Jeffrey. Actually, oscillatory and exponential approximation formulas, obtained far from turning points, were independently used by Green (1837) and Liouville (1837) and they can be already found in an investigation on the motion of a planet in an unperturbed elliptic orbit by the Italian astronomer Carlini (1817). For historical notes on the WKB-method we refer to [\[7,11,23\]](#).

The WKB method can be also applied to three-dimensional problems only under some particular circumstances; for instance when the potential is spherically symmetric and the radial differential equation can be separated.

In general WKB approximation is not suitable for problems in dimension  $n$  higher than 1. Actually, semiclassical methods in higher dimension require new sophisticated tools such as the *Agmon metric*, *microlocal calculus* and  *$\hbar$ -pseudodifferential operators*. These tools have been developed by Agmon, Hörmander and Maslov in the 1970s and since then a large number of mathematicians have contributed to this subject. In Sects. [“Semiclassical Approximation in Any Dimension”](#) and [“Propagation of Quantum Observables”](#) we briefly introduce the reader to the basic concepts of these theories and we resume the most important results such as the exponential decay of

wave-functions and the Egorov Theorem, which asymptotically describes the quantum evolution of an observable by means of the classical evolution of its classical counterpart.

Actually, these methods may be applied to many other fields, where  $\hbar$  is a small quantity not related to the Planck constant but it may represent a different small physical quantity such as the adiabatic parameter in adiabatic problems, the (square root of the) inverse of the heavy mass in the Born–Oppenheimer approximation, etc.

## Notation

Hereafter, for the sake of definiteness, let us fix  $2m = 1$ .

## The WKB Approximation

If the potential  $V(x)$  does not have a very simple form then the solution of the time-independent Schrödinger equation even in one dimension

$$-\hbar^2\psi'' + V(x)\psi = E\psi, \quad x \in (x_1, x_2), \quad (3)$$

is a quite complicated problem which requires the use of a sort of approximation method (hereafter  $' = d/dx$  denotes the derivative with respect  $x$ ); here  $(x_1, x_2)$  is a given finite or infinite one-dimensional interval.

We distinguish different regions: the classically allowed regions where  $V(x) < E$  and the classically forbidden regions where  $V(x) > E$ . Approximation formulas for the solution of Eq. (3) in these separate regions have been studied since Carlini, Green, Liouville and Jacobi. The problem to connect these approximated solutions across the turning points was raised in the framework of Quantum Mechanics and it was independently solved by Jensen, Wentzel, Kramers and Brillouin. For the general treatment of semiclassical approximation in dimension one and for physical applications we refer to [\[2,3,10,21,24,25,33\]](#).

## Semiclassical Solutions

### in the Classically Allowed Region

The basic idea is quite simple: if  $V = \text{const.}$  is a constant smaller than  $E$  then Eq. (3) has solutions  $e^{\pm ikx/\hbar}$  for a suitable real-valued constant  $k$ . This fact suggests to us that if the potential, while no longer constant, varies only slowly with  $x$  and it is such that  $V(x) < E$  for any  $x \in (x_1, x_2)$ , we might try a solution of the form

$$\psi(x) = a(x)e^{iS(x)/\hbar}, \quad (4)$$

except that the *amplitude*  $a(x)$  is not constant and the *phase*  $S(x)$  is not simply proportional to  $x$ . Substituting (4)

into (3) and separating the real and imaginary parts of the coefficients of  $e^{iS(x)/\hbar}$ , then we get a system of equations for the  $x$ -dependent real-valued phase  $S(x)$  and amplitude  $a(x)$ :

$$\begin{cases} u^2 - p^2(x) = \hbar^2 a''/a \\ (a^2 u)' = 0 \end{cases}$$

where we set

$$u(x) = S'(x) \quad \text{and} \quad p(x) = \sqrt{E - V(x)}.$$

Hence,  $a(x) = C[u(x)]^{-1/2}$  for some constant  $C$ .

Since the complex conjugation  $a(x)e^{-iS(x)/\hbar}$  of (4) is a solution of the same equation and the Wronskian between these two solutions is not zero then the general solution of Eq. (3) can be written as

$$\begin{aligned} \psi(x) &= b_+ \psi_+(x) + b_- \psi_-(x), \\ \psi_{\pm}(x) &= \frac{1}{\sqrt{u(x)}} e^{\pm i/\hbar \int u(x) dx}. \end{aligned} \quad (5)$$

Here,  $b_{\pm}$  are arbitrary constants and  $u = u(x; \hbar)$  is a real-valued solution, depending on the variable  $x$  and on the semiclassical parameter  $\hbar$ , of the nonlinear ordinary differential equation

$$\mathcal{F}(u) = u^2 - p^2 - \hbar^2 u f(u) = 0 \quad (6)$$

where

$$f(u) = u^{-1/2} (u^{-1/2})''.$$

The fact that (6) is a nonlinear equation, whereas the Schrödinger equation (3) is linear, would be usually regarded as a drawback, but we shall take advantage of the nonlinearity to develop a simple approximation method for solving (6).

Taking the limit of  $\hbar$  small in (6) then it turns out that the leading order of  $u$  is simply given by  $p$ . Actually, it is possible to prove that (3) has twice continuously differentiable solutions  $\psi_{\pm}$  satisfying the following asymptotic behavior

$$\begin{aligned} \psi_{\pm}(x) &= \frac{1}{\sqrt{p(x)}} e^{\pm i/\hbar \int_a^x p(q) dq} [1 + \delta_{\pm}(x)], \\ x &\in (x_1, x_2), \end{aligned} \quad (7)$$

where  $a$  is an arbitrary point in  $(x_1, x_2)$  and where

$$|\delta_{\pm}(x)| \leq e^{\frac{1}{2}\hbar \left| \int_a^x |f[p(q)]| dq \right|} - 1 = \mathcal{O}(\hbar).$$

Henceforth, the dominant term of the solution (5) has an oscillating behavior of the form

$$\begin{aligned} \psi(x) &= b_+ \frac{e^{i/\hbar \int_a^x \sqrt{E-V(q)} dq}}{\sqrt[4]{E-V(x)}} \\ &\quad + b_- \frac{e^{-i/\hbar \int_a^x \sqrt{E-V(q)} dq}}{\sqrt[4]{E-V(x)}} + \mathcal{O}(\hbar). \end{aligned}$$

In order to compute also the other terms of the asymptotic expansion of the solutions  $\psi_{\pm}(x)$  we formally solve the nonlinear Eq. (6) by means of the formal power series in  $\hbar^2$ :

$$u = u(x, \hbar) = \sum_{n=0}^{\infty} \hbar^{2n} u_{2n}(x). \quad (8)$$

The functions  $u_{2n}(x)$  are determined explicitly, order by order, by formally substituting (8) into (6) and requiring that the coefficients of the same terms  $\hbar^{2n}$  are zero for any  $n \geq 0$ . In such a way and assuming that the potential  $V(x)$  admits derivatives at any order then we obtain that

$$\begin{aligned} u_0(x) &= p(x) \\ u_2(x) &= -\frac{1}{4} \frac{p''(x)}{p^2(x)} + \frac{3}{8} \frac{[p'(x)]^2}{p^3(x)} \\ u_4(x) &= \frac{1}{16} \frac{p''''(x)}{p^4(x)} - \frac{5}{8} \frac{p'''(x)p'(x)}{p^5(x)} - \frac{13}{32} \frac{[p''(x)]^2}{p^5(x)} \\ &\quad + \frac{99}{32} \frac{p''(x)[p'(x)]^2}{p^6(x)} - \frac{297}{128} \frac{[p'(x)]^4}{p^7(x)}, \end{aligned}$$

and so on. Thus, the asymptotic behavior (7) can be improved up to any order. In particular, let

$$\psi_{N,\pm}(x) = \frac{1}{\sqrt{p_N(x)}} e^{\pm i/\hbar \int_a^x p_N(q) dq}, \quad x \in (x_1, x_2),$$

where  $a \in (x_1, x_2)$  is arbitrary and fixed and  $p_N(x) = \sum_{n=0}^N u_{2n}(x) \hbar^{2n}$ , in particular  $p_0(x) = p(x)$ , let us introduce the *error-control function*

$$\epsilon_N(x) = \frac{1}{\hbar p_N(x)} \mathcal{F}[p_N(x)] = \mathcal{O}(\hbar^{2N+1})$$

where  $\epsilon_0(x) = \hbar f[p(x)]$ ; then the two functions  $\psi_{N,\pm}(x)$  approximate the solutions  $\psi_{\pm}(x)$  of Eq. (3) up to the order  $2N + 1$ , that is

$$\begin{aligned} |\psi_{\pm}(x) - \psi_{N,\pm}(x)| &\leq \left[ \exp \left( \frac{1}{2} \left| \int_a^x |\epsilon_N(q)| dq \right| \right) - 1 \right] \\ &= \mathcal{O}(\hbar^{2N+1}), \quad \forall x \in (x_1, x_2). \end{aligned}$$

We underline that this result is valid whether or not  $x_1$  and  $x_2$  are finite, also whether or not  $V(x)$  is bounded: it suffices that the *error-control function* is absolutely integrable:  $\epsilon_N \in L^1(x_1, x_2)$ .

### Semiclassical Solutions in the Classically Forbidden Region

The previous asymptotic computation of the solutions  $\psi_{\pm}$  applies also in the classically forbidden region where  $V(x) > E$ , provided that  $p(x)$  is one of the two purely imaginary determination of  $\sqrt{E - V(x)}$ . By assuming that  $\Im p(x) < 0$  then Eq. (3) has twice continuously differentiable solutions of the form

$$\psi_{\pm}(x) = \frac{1}{\sqrt{|p(x)|}} e^{\pm i/\hbar \int_a^x |p(q)| dq} [1 + \delta_{\pm}(x)], \quad x \in (x_1, x_2), \quad (9)$$

where  $a$  is an arbitrary point in  $(x_1, x_2)$  and where the remainder terms  $\delta_{\pm}$  are bounded as follows

$$|\delta_{\pm}(x)| \leq e^{\frac{1}{2}\hbar \left| \int_a^x |f[p(q)]| dq \right|} - 1 = \mathcal{O}(\hbar)$$

where  $a_+ = x_1$  and  $a_- = x_2$ . In particular, if  $V(x) > E$  for any  $x \in (-\infty, x_2)$ , that is  $x_1 = -\infty$  (and similarly we can consider the case where  $x_2 = +\infty$ ), and the error-control function  $f[p(x)] \in L^1(-\infty, x_2)$  then the above asymptotic estimate holds for any  $x \in (-\infty, x_2)$ . We also emphasize that the asymptotic behavior (9) could be also improved up to any order, as done for the approximated solutions in the classically allowed regions.

We underline that the two solutions  $\psi_{\pm}$  play here a different role. Indeed, solution  $\psi_{-}(x)$  is an exponentially decreasing function as  $x > a$  grows, while  $\psi_{+}(x)$  is exponentially increasing. The first solution is usually called *recessive* solution and it is *uniquely determined* by the asymptotic power expansion as  $\hbar \rightarrow 0$ . The other solution is called *dominant* solution and it is *not uniquely determined* by the asymptotic expansion; in fact,  $\psi_{+}(x) + c\psi_{-}(x)$ ,  $x > a$ , is still a solution of Eq. (3) which has the same asymptotic behavior of  $\psi_{+}(x)$  for any  $c \in \mathbb{C}$ .

### Connection Formula

It is not possible to use the semiclassical solutions (7) and (9) at turning points  $x_E$  since the term  $1/p$  diverges and there is no guarantee that the same combination of simple semiclassical solutions will fit the same particular solution on both sides of the turning point. This is the so-called *connection problem*: if the values  $b_{\pm}$  of the semiclassical

solution are known in a given region then the problem consists of computing the values of  $b_{\pm}$  in a different region separated from the first region by one (or more) turning points.

The main two methods used to treat this problem are the following ones:

- The *complex method*, where the two regions surrounding the turning point are joined by a path in the complex plane which is sufficiently far from the turning points. In such a case the semiclassical solutions in different regions are connected by means of holomorphic extension arguments. Thus, in order to apply this method we have to assume that the potential  $V(x)$  can be holomorphically extended to the complex plane.
- The second method employs the technique of the *uniform approximation*, where the required solution is mapped on the solution of a simpler and suitable equation which has the same disposition of turning points as the original one. As well as enabling the connection problem to be solved, this method also provides the wave function in the neighborhood of the turning points, which was bypassed in the complex method.

We are going now to explain these methods in detail.

**The Complex Method** By assuming that  $V(x)$  is the restriction on the real axis of a holomorphic function  $V(z)$ ,  $z \in \mathbb{C}$ , we turn now to the approximate solution of the differential equation

$$-\hbar^2 \frac{d^2 \psi(z)}{dz^2} = [E - V(z)] \psi(z)$$

in a complex domain  $\mathcal{D}$  in which  $V(z)$  is holomorphic and  $p^2(z) = E - V(z)$  does not vanish. Then, this equation has two holomorphic solutions

$$\psi_{\pm}(z) = \frac{1}{\sqrt{p(z)}} e^{\pm i/\hbar \int_a^z p(q) dq} [1 + \mathcal{O}(\hbar)], \quad z \in \mathcal{D}, \quad (10)$$

where  $a$  is an arbitrary point; in particular, for our purposes it is convenient to choose it coinciding with the turning point, i. e.:  $a = x_E$ . Actually, this asymptotic behavior holds under a technical assumption: given a reference point  $a_+$  (respectively  $a_-$ ) then (10) is true for  $\psi_{+}(z)$  (resp.  $\psi_{-}(z)$ ) for any  $z$  such that there exists a *progressive* (resp. *regressive*) path  $\Gamma_+$  (resp.  $\Gamma_-$ ) contained in  $\mathcal{D}$  and connecting  $z$  with  $a_+$  (resp.  $a_-$ ); that is as  $q$  passes along  $\Gamma_+$  (resp.  $\Gamma_-$ ) from  $a_+$  (resp.  $a_-$ ) to  $z$  then  $\Re[\xi(z)]$  is non-

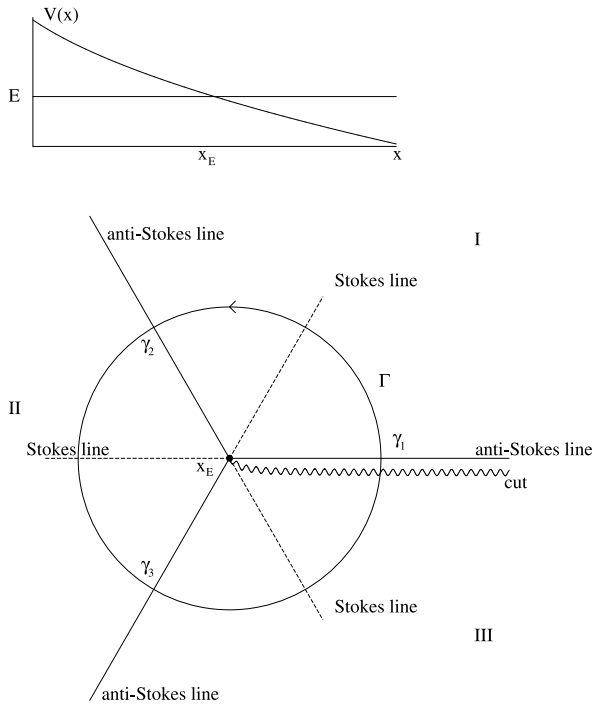


decreasing (resp. nonincreasing) where

$$\xi(z) = \frac{i}{\hbar} \int_{x_E}^z p(q) dq.$$

Now, we denote as a *Stokes line* any path coming from the turning point  $x_E$  such that  $\Im[\xi(z)] = 0$ ; we denote as an *anti-Stokes line*, or *principal line*, any path coming from the turning point  $x_E$  such that  $\Re[\xi(z)] = 0$ . Classically forbidden (resp. allowed) regions are particular cases of Stokes (resp. anti-Stokes) lines, which lie on the real axis.

Let us consider now the case, as in Fig. 1, where the turning point  $x_E$  is simple (that is  $V'(x_E) \neq 0$ ),  $p^2(x) < 0$  for  $x < x_E$  and  $p^2(x) > 0$  for  $x > x_E$ . Since the turning point  $x_E$  is simple then in a neighborhood of the turning point we have that  $p(z) \approx c[z - x_E]^{\frac{1}{2}}$ , for some  $c \in \mathbb{R}^+$ , and  $\xi(z) \approx \frac{2ic}{3\hbar}[z - x_E]^{\frac{3}{2}}$ . Thus, Stokes lines are three different paths coming from the turning point and with asymptotic directions  $\arg(z - x_E) = \frac{1}{3}\pi, \pi, \frac{5}{3}\pi$ ; while the asymptotic directions of the anti-Stokes lines are given by  $\arg(z - x_E) = 0, \frac{2}{3}\pi, \frac{4}{3}\pi$ .



**Perturbation Theory, Semiclassical, Figure 1**

Stokes and anti-Stokes lines in a neighborhood of a simple turning point  $x_E$  where the left-hand side of the turning point is classically forbidden and the right-hand side is classically allowed. The wavy line denotes the cut of the multi-valued function  $p(z)$

Along the three anti-Stokes lines  $\gamma_1, \gamma_2$  and  $\gamma_3$  the solution  $\psi$  can be written as

$$\psi(z) = b_{+,j}\psi_{+,j}(z) + b_{-,j}\psi_{-,j}(z), \quad z \in \gamma_j, \quad j = 1, 2, 3,$$

where  $\psi_{\pm,j}$  have the asymptotic behavior given by (10) and where the coefficients  $b_{\pm,j}$  are asymptotically fixed along the anti-Stokes lines. Let  $F_j$  be the  $2 \times 2$  matrix which connects the coefficients:

$$\begin{pmatrix} b_{+,j+1} \\ b_{-,j+1} \end{pmatrix} = F_j \begin{pmatrix} b_{+,j} \\ b_{-,j} \end{pmatrix}.$$

In the region I, with boundaries given by the anti-Stokes lines  $\gamma_1$  and  $\gamma_2$ , the solution  $\psi_+$  is *dominant* and thus the coefficient  $b_+$  cannot change in this region, similarly in the region III with boundaries  $\gamma_3$  and the cut. In contrast, in the region II, with boundaries given by the anti-Stokes lines  $\gamma_2$  and  $\gamma_3$ , the solution  $\psi_-$  is *dominant* and thus the coefficient  $b_-$  cannot change in this region. Then the matrices  $F_j$  can be written as

$$F_1 = \begin{pmatrix} 1 & 0 \\ r & 1 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$$

$$\text{and } F_3 = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

where  $r, s$  and  $t$  have to be determined. To this end let us consider the closed anti-clockwise path  $\Gamma$  surrounding the simple turning point  $x_E$ ; along this path the exponential term gains a phase  $v = \frac{1}{\hbar} \oint_{\Gamma} p(z) dz$  and the argument of  $p(z)$ , which appears in (10), decreases of  $\pi$ . Hence, the holomorphic extension of  $\psi_{\pm}$ , around this closed path, are thus proportional to the solution  $\psi_{\mp}$ :

$$\psi_+ \rightarrow ie^{iv}\psi_- \quad \text{and} \quad \psi_- \rightarrow ie^{-iv}\psi_+.$$

Therefore, the numbers  $r, s$  and  $t$  are such that

$$F_1 F_2 F_3 = \begin{pmatrix} 1 + st & s \\ r + (rs + 1)t & rs + 1 \end{pmatrix} \sim \begin{pmatrix} 0 & ie^{-iv} \\ ie^{iv} & 0 \end{pmatrix}$$

from which it follows that

$$s = ie^{-iv}, \quad r = t = ie^{iv}.$$

In particular, since  $\int p(z) dz$  does not diverge at  $x_E$  then we can take  $\Gamma$  in a small neighborhood of  $x_E$  and  $v \sim 0$ . Thus, the Stokes' rule follows: the connection between the two coefficients from one Stokes line to the (anti-clockwise) adjacent one follows the following rule

$$\begin{aligned} b_{\text{dominant}} &\rightarrow b_{\text{dominant}} \quad \text{and} \\ b_{\text{recessive}} &\rightarrow b_{\text{recessive}} + ib_{\text{dominant}}. \end{aligned}$$

**The Method of Comparison Equations** With this method we obtain a local approximate solution, in a neighborhood of the turning points, in terms of the known solutions of an equivalent equation

$$\frac{d^2 \phi}{dy^2} + q(y)\phi(y) = 0, \quad (11)$$

where  $q(y)$  is chosen to be similar in some way to  $p^2(x)$ , but simpler in order to have an explicit solution. For the sake of definiteness, we restrict here our analysis to the case of a simple turning point  $x_E$ .

In order to get an approximate solution to Eq. (3) in a small neighborhood of a simple zero  $x_E$  of  $p^2(x)$  we introduce the following changes of variable

$$x \rightarrow y(x) = \left[ \frac{3}{2} \xi(x) \right]^{\frac{2}{3}}, \quad \xi(x) = \frac{i}{\hbar} \int_{x_E}^x p(q) dq.$$

Equation (3) takes the form of the approximate Eq. (11) where  $q(y) \approx -y$  in a neighborhood of the turning point, which admits solutions given by the Airy functions  $Ai(y)$  and  $Bi(y)$ . Then, the approximate solution of the Schrödinger equation (3) in a neighborhood of the turning point has the form

$$\psi(x) \sim \alpha \left[ \frac{\pi^2 y(x)}{p^2(x)} \right]^{\frac{1}{4}} \{ \cos(\mu) Ai[y(x)] + \sin(\mu) Bi[y(x)] \}$$

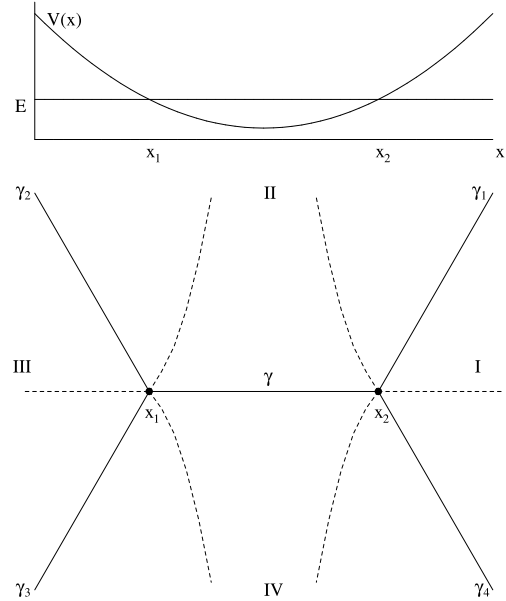
where  $\alpha$  and  $\beta$  are constants. The Airy functions are well understood and exact connection formulas can be established obtaining that

$$\psi(x) \rightarrow \frac{\alpha}{2\sqrt{|p(x)|}} \left[ \cos(\mu) e^{-|\xi|} + 2 \sin(\mu) e^{|\xi|} \right]$$

from the turning point to the classically forbidden region, and

$$\psi(x) \rightarrow \frac{\alpha}{\sqrt{|p(x)|}} \cos \left( |\xi| + \mu - \frac{1}{4}\pi \right)$$

from the turning points to the classically allowed region. Thus, we have established a connection between oscillatory and exponentially increasing and decreasing solutions. Clearly, this entire approach breaks down if the energy is too close to a value corresponding to a stationary point of the potential. In this case a different approximation must be implemented. For instance, in the case of a turning point with multiplicity 2 then the approximate solution of the Schrödinger equation (3) is given by means of Weber parabolic cylinder functions.



**Perturbation Theory, Semiclassical, Figure 2**

Stokes and anti-Stokes lines in a neighborhood of the bottom of a single well

### Bound States for a Single Well Potential

We apply now the previous techniques in order to compute the stable states, that is normalized solutions of the Eq. (2), when the potential  $V(x)$  has a single well shape; that is it has a simple minimum point  $x_{\min}$ , with minimum value  $V_{\min}$  such that  $V(x) > V_{\min}$  for any  $x \neq x_{\min}$  and  $\liminf_{|x| \rightarrow +\infty} V(x) > V_{\min}$ . Then, for  $E > V_{\min}$  close enough to the bottom of the well, equation  $p(x) = 0$  has only two simple solutions  $x_1 < x_{\min} < x_2$  and the typical picture of the Stokes lines appears as in Fig. 2. Here,  $\psi_{\pm, x_1}$  (resp.  $\psi_{\pm, x_2}$ ) denotes the fundamental solutions with asymptotic behaviors (7) and (9) in a neighborhood of  $x_1$  (resp.  $x_2$ ). Since we are looking for normalized solutions then in the classically forbidden region  $x > x_2$ , contained in the region I, the *dominant* term  $\psi_{+, x_2}$  of the solution  $\psi$  should have coefficient  $b_{+, x_2}^I$  exactly zero, that is

$$\psi(z) = \psi_{-, x_2}(z), \quad z \in I,$$

where the coefficients  $b_{-, x_2}$  of the *recessive* solution  $\psi_{-, x_2}$  is chosen equal to 1. Turning around the turning point  $x_2$  and crossing the anti-Stokes line it follows that the coefficients of the solution  $\psi = b_{+, x_2}^{II} \psi_{+, x_2}(x) + b_{-, x_2}^{II} \psi_{-, x_2}(x)$  in the region II take the form

$$\begin{pmatrix} b_{+, x_2}^{II} \\ b_{-, x_2}^{II} \end{pmatrix} = F_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

where

$$F_2 \sim \begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix}$$

according to the Stokes' rule.

In order to study the matching condition around the other turning point  $x_1$  we have to transport the origin of integration at  $x_1$  obtaining

$$\psi(z) = b_{+,x_1}^{\text{II}} \psi_{+,x_1}(z) + b_{-,x_1}^{\text{II}} \psi_{-,x_1}(z)$$

where

$$b_{+,x_1}^{\text{II}} = \psi_{+,x_2}(x_1) b_{+,x_2}^{\text{II}} \quad \text{and} \quad b_{-,x_1}^{\text{II}} = \psi_{-,x_2}(x_1) b_{-,x_2}^{\text{II}}.$$

On the other hand, also in region III the *dominant* term  $\psi_{+,x_1}$  of the solution  $\psi$  should have coefficient  $b_{+,x_1}^{\text{III}}$  exactly zero; from this fact and from the Stokes' rule applied to the turning point  $x_1$  it follows that the equation for the dominant term of the bound states is

$$\psi_{+,x_2}(x_1) + \psi_{-,x_2}(x_1) = 0,$$

which implies

$$\cos \left[ \frac{1}{\hbar} \left( \int_{x_1}^{x_2} p(x) dx + \mathcal{O}(\hbar) \right) \right] = 0.$$

Hence, we obtain the well known Bohr–Sommerfeld rule

$$\int_{x_1}^{x_2} p(x) dx = \hbar \left[ \frac{1}{2} \pi + n\pi \right] + \mathcal{O}(\hbar^2), \quad n \in \mathbb{N}.$$

### Double Well Model: Estimate of the Splitting and the “Flea of the Elephant”

We consider now the case of a symmetric double well potential; that is  $V(-x) = V(x)$  and it has two simple minima at  $x_+ > 0$  and  $x_- = -x_+$  separated by a barrier, we assume also that the minimum value  $V_{\min} < \liminf_{|x| \rightarrow \infty} V(x)$ . For instance,  $V(x) = x^4 - 2\alpha x^2$  for some  $\alpha > 0$ ; in this case the shape of the potential has two symmetric wells where  $x_{\pm} = \pm \sqrt{\alpha}$  and  $V_{\min} = -\alpha^2 < 0$ , the two wells are separated by an energy barrier with top  $V_{\max} = 0$  at  $x = 0$ . The semiclassical double well model is not only a very enlightening pedagogical problem, but it is also the basic argument explaining many relevant physical questions, see, e. g., [5,12,16,32,34].

In the interval  $(V_{\min}, V_M)$ , where

$$V_M = \min \left[ V_{\max}, \liminf_{|x| \rightarrow \infty} V(x) \right],$$

the stable states appear as *doublets*  $E_{\pm}$  whose distance  $\omega$ , named *splitting*, is quite small. The associated eigenvectors are even and odd (real-valued) wave-functions, that is

$$\psi_{\pm}(-x) = \pm \psi_{\pm}(x). \quad (12)$$

The splitting  $\omega = E_- - E_+$  can be computed as

$$\omega = \hbar^2 \frac{\psi_+(0) \psi'_-(0)}{\int_0^{\infty} \psi_-(x) \psi_+(x) dx}$$

and it turns out to be exponentially small as  $\hbar \rightarrow 0$ . More precisely, if  $E_+$  is the ground state then  $\omega = \mathcal{O}(e^{-S_0/\hbar})$  where

$$S_0 = \int_{x_-}^{x_+} \sqrt{V(x) - V_{\min}} dx$$

is the Agmon distance between the two wells.

We would emphasize that the eigenvectors  $\psi_{\pm}$  are asymptotically localized on both wells because of the symmetry property (12). The effect on the ground state of a small perturbation  $W(x)$  that breaks the symmetry is worth mentioning. In such a case the property (12) does not work and, even if the small perturbation  $W(x)$  is supported only on one side and far from the bottom of the well, the ground state, instead of being asymptotically supported on both wells, may be localized on just one well. According to Barry Simon we may state that *the perturbation  $W(x)$  is a small flea on the elephant  $V(x)$ . The flea does not change the shape of the elephant* – in the sense that the splitting is still exponentially small – *but it can irritate the elephant enough so that it shifts its weight* – in the sense that the ground state is localized on just one well.

### Semiclassical Approximation in Any Dimension

Here we consider the eigenvalue problem (2) in any dimension  $n \geq 1$  where the potential  $V$  is assumed to be a *multi-well potential* [6,17,18,31]. More precisely, we assume that

$$V_{\min} := \inf V(x) < V_{\infty} = \liminf_{|x| \rightarrow +\infty} V(x)$$

and for any  $E \in (V_{\min}, V_{\infty})$  we can write the decomposition of

$$V^{-1}((-\infty, E]) = \cup_{j=1}^N U_j$$

as the union of  $N$  disjoint, compact and connected sets  $U_j$ . Inside these sets, named *wells*, the classical motion is allowed; outside the classical motion is forbidden.

From well-known results we have that for energies in the interval  $(V_{\min}, V_{\infty})$  the eigenvalue problem (2) admits only discrete spectrum, that is we have only isolated eigenvalues with finite multiplicity. In order to compute these eigenvalues we'll consider, at first, the solutions of Eq. (2) inside any single well and then we'll take into account the tunneling effect among the adjacent wells as a perturbation.

Actually, it is also possible to consider the case where  $E > V_{\infty}$ . In such a case we don't expect to have isolated eigenvalues but, under some circumstances, resonant states [13,19,20]. However, we don't fix our attention here on this problem.

### Semiclassical Eigenvalues at the Bottom of a Well

Let  $x_0$  be a local nondegenerate minimum for the potential  $V(x)$ . For the sake of definiteness we assume that  $x_0 = 0$  and the local minimum is such that  $V(0) = \nabla V(0) = 0$  and the Hessian matrix  $\text{Hess} = (\partial^2 V(0)/\partial x_i \partial x_j)_{i,j=1\dots n}$  is positively defined with eigenvalues  $2\lambda_j > 0$ ,  $j = 1, \dots, n$ ; furthermore, we choose the system of coordinates such that the Hessian matrix has diagonal form, i. e.  $\text{Hess} = \text{diag}(2\lambda_j)$ .

In order to compute the eigenvalues at the bottom of the well containing the minimum  $x_0$  we approximate the potential by means of the  $n$ -dimensional harmonic oscillator. Thus, Eq. (2) takes the form

$$\sum_{j=1}^n \left[ -\hbar^2 \frac{\partial^2 \psi}{\partial x_j^2} + \lambda_j x_j^2 - E \right] \psi = 0$$

which has exact eigenvalues

$$\hbar \left[ \sum_{j=1}^n \sqrt{\lambda_j} (2\alpha_j + 1) \right], \quad \alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n,$$

with normalized eigenvectors

$$(\hbar\pi)^{-n/4} \prod_{j=1}^n (2^{\alpha_j} \alpha_j!)^{-1/2} \lambda_j^{1/8} \cdot e^{-\sqrt{\lambda_j} x_j^2 / 2\hbar} H_{\alpha_j} \left( \lambda_j^{1/4} \hbar^{-1/2} x_j \right),$$

where  $H_m$  is the Hermite polynomial of degree  $m$ .

It is clear that when  $\hbar$  is small enough then the first energy level of the harmonic oscillator is very close to the bottom of the potential and thus we expect that such an approximation gives the leading term of the first energy

level and wave-function of (2). In fact, there exist two formal power series

$$E(\hbar) \sim \sum_{j=0}^{\infty} \hbar^j e_j \quad \text{and} \quad a(x; \hbar) \sim \sum_{r=0}^{\infty} \hbar^r a_r(x), \quad (13)$$

where  $e_0 = V_{\min} = 0$ ,  $\hbar e_1 = \hbar \sum_{j=1}^n \sqrt{\lambda_j}$  is the first eigenvalue of the associated harmonic oscillator and  $a_0 = [\prod_{j=1}^n \lambda_j / \pi^2]^{1/8}$ , such that the function

$$\psi(x; \hbar) = \hbar^{-n/4} a(x; \hbar) e^{-\varphi(x)/\hbar} \quad (14)$$

is such that

$$-\hbar^2 \Delta \psi + [V(x) - E(\hbar)] \psi = \mathcal{O}(\hbar^\infty) e^{-\varphi(x)/\hbar}$$

in a neighborhood of  $x_0 = 0$ , where  $\varphi(x) = \frac{1}{2} \sum_{j=1}^n \sqrt{\lambda_j} x_j^2 + \mathcal{O}(|x|^2)$  is the positive solution of the Eikonal equation

$$|\nabla \varphi|^2 = V - e_0$$

in a neighborhood of  $x_0 = 0$ . The way to obtain this result essentially consists of inserting the formal power series (14) in Eq. (2) and expanding in powers of  $\hbar$  the coefficients of  $e^{-\varphi(x)/\hbar}$ , then of requiring the cancelation of the coefficients of  $\hbar^j$  for any  $j \in \mathbb{N}$ .

These approximate solutions are valid in a neighborhood of the nondegenerate minimum and they are the natural candidates to become the true eigenvalue and eigenfunction of the problem (2). In fact, we'll extend this single well approximate solution (14) to a larger domain and then we'll solve the *connection* problem.

To this end we fix an open, sufficiently small, regular bounded set  $\Omega$  containing the minimum point  $x_0$ ; then Eq. (2) with Dirichlet boundary condition on  $\Omega$ , that is

$$\psi|_{\partial\Omega} = 0, \quad \|\psi\|_{L^2(\Omega)} = 1, \quad (15)$$

admits one simple eigenvalue  $E_\Omega(\hbar)$  at the bottom of the well, close to the first eigenvalue of the harmonic oscillator  $e_0 + \hbar e_1$ , which admits the asymptotic expansion (13).

Actually, such an eigenvalue will depend on the choice of the domain  $\Omega$ . More precisely, if  $E_{\Omega'}(\hbar)$  denotes the first eigenvalue of Eq. (2) with Dirichlet boundary condition on  $\Omega'$  for a different domain  $\Omega'$  containing  $x_0$ , then  $E_{\Omega'}(\hbar)$  differs from  $E_\Omega(\hbar)$  for an exponentially small term. Furthermore, it is also possible to prove that, modulo an exponentially small error, the spectrum of the Schrödinger equation (2) in some interval depending on  $\hbar$  is the same as the spectrum of the direct sum of the Dirichlet single well problems in the same interval.

### Agmon Metric

In order to study the exponential behavior of the solution (14) when  $x$  is far from the minimum  $x_0$  we introduce now the Agmon distance between two points  $x_1, x_2 \in \mathbb{R}^n$  defined as

$$d_E(x_1, x_2) = \inf_{\gamma} \int_{\gamma} \sqrt{[V(x) - E]_+} dx,$$

where

$$[V(x) - E]_+ = \max\{V(x) - E, 0\},$$

$\gamma$  is any regular path connecting the two points  $x_1$  and  $x_2$  and  $E \geq V_{\min}$  is fixed. We underline that the Agmon distance defines a (pseudo)-metric on the Euclidean space  $\mathbb{R}^n$ . Indeed, it satisfies the triangle inequality

$$d_E(x_1, x_2) \leq d_E(x_1, x_3) + d_E(x_3, x_2), \quad \forall x_1, x_2, x_3 \in \mathbb{R}^n,$$

but it is degenerate on each well  $U$  where  $V < E$ , in particular  $d_E(x_1, x_2) = 0$  for any couple of points  $x_1$  and  $x_2$  belonging to the same well and the *diameter* of each well  $U$  is zero with respect to this metric.

This kind of (pseudo)-metric has been used to obtain precise exponential decay of wave-functions of Schrödinger operators. Indeed, the exponential term  $\varphi(x)$  in (14) is simply given by

$$\varphi(x) = d_{e_0}(x_0, x)$$

for any  $x$  in a neighborhood of  $x_0$ .

Furthermore, by means of the Agmon metric we are also able to give an estimate of the exponential decay of wavefunctions not only for  $x$  in a neighborhood of the minimum  $x_0$ , but also for  $x$  far from this point. In particular, let  $V_{\min} < E < V_{\infty}$  be fixed and let  $U$  be one of the wells, we consider now the eigenvalue problem for the Schrödinger equation with Dirichlet boundary conditions on a regular open set  $\Omega$  containing the well  $U$ . Let  $E_{\Omega} < E$  be an eigenvalue of Eq. (2) with Dirichlet boundary condition (15) with associated eigenfunction  $\psi_{\Omega}$ . Then, the Agmon theorem enables us to obtain the following decay estimate for the eigenfunction  $\psi_{\Omega}$ : for any fixed and positive  $\epsilon$  then

$$\left\| \nabla \left[ e^{d_E(x, U)/\hbar} \psi_{\Omega} \right] \right\|_{L^2(\Omega)} + \left\| e^{d_E(x, U)/\hbar} \psi_{\Omega} \right\|_{L^2(\Omega)} \leq C_{\epsilon} e^{\epsilon/\hbar}$$

for some positive constant  $C_{\epsilon}$  depending on  $\epsilon$  (of course we expect that  $C_{\epsilon}$  will grow as  $\epsilon$  goes to zero). That is we

have a good a priori estimate of the wavefunction  $\psi_{\Omega}$  in the weighted  $H^1(\Omega)$  space with weight  $e^{d_E(x, U)/\hbar}$ . Since  $d(x, U) = 0$  for any  $x$  belonging to the well  $U$  then it follows that the solution  $\psi_{\Omega}$  is exponentially decreasing outside the classical allowed region, as already seen in the one-dimensional problems.

### Tunneling Between Wells

In order to consider the tunneling effect among the wells  $\{U_j\}_{j=1}^N$  for any fixed  $E \in (V_{\min}, V_{\infty})$  we define

$$d_E(U_i, U_j) = \inf_{x \in U_i, y \in U_j} d_E(x, y)$$

as the Agmon distance between the two wells  $U_i$  and  $U_j$ , by construction it turns out that this distance is strictly positive for any  $i \neq j$ . A special role will be played by the minimal distance among these wells

$$S_0 = \min_{i \neq j} d_E(U_i, U_j).$$

In fact, the discrete spectrum of the Dirichlet realization of the Schrödinger equation on the boundary of the wells give, up to error of the order  $e^{-S_0/\hbar}$ , the discrete spectrum of the original Schrödinger equation (2). In order to be more definite, let  $M_j^{S, \eta}$  be the open set containing the well  $U_j$  defined as

$$M_j^{S, \eta} = \{x \in \mathbb{R}^n : d_E(x, U_j) < S \text{ and } d_E(x, U_k) > \eta, k \neq j\}$$

that is  $M_j^{S, \eta}$  is, essentially, a ball (with respect to the Agmon pseudo-metric) large enough centered in the well  $U_j$  where we have eliminated the points from the other wells. If necessary we can regularize the boundary of  $M_j^{S, \eta}$ . In the following we take  $\eta > 0$  small enough and  $S$  large enough, in particular we require that  $S > 2S_0$ . Let  $\sigma_j$  be the discrete spectrum of the Schrödinger equation (2) with Dirichlet condition on  $M_j^{S, \eta}$  and let  $\sigma$  be the discrete spectrum of the Schrödinger equation (2). Then, for any interval  $I(\hbar) = [\alpha(\hbar), \beta(\hbar)]$ , where  $\alpha(\hbar), \beta(\hbar) \rightarrow e_0$  as  $\hbar \rightarrow 0$ , there exists a bijection

$$b: \sigma \cap I(\hbar) \rightarrow \left[ \bigcup_{j=1}^N \sigma_j \right] \cap I(\hbar)$$

such that for any  $\rho < S_0 - 2\eta$

$$|b(\lambda) - \lambda| \leq C_{\rho} e^{-\rho/\hbar} \quad (16)$$

for some  $C_{\rho} > 0$ .



Actually, this result could be improved by assuming some regularity properties on the potential  $V$  and estimate (16) can be replaced by the precise asymptotic behavior.

If we consider now the *symmetric double well* problem where  $N = 2$  and with symmetric potential, e.g.  $V(x_1, x_2, \dots, x_n) = V(-x_1, x_2, \dots, x_n)$ , then, by symmetry, we have that the two Dirichlet realizations coincide, up to the inversion  $x_1 \rightarrow -x_1$ . Then  $\sigma_1 = \sigma_2$ . If we denote by  $E_1$  and  $E_2$  the first two eigenvalues of  $\sigma$  then  $E_1 < E_2$  (in fact, the first eigenvalue is always nondegenerate) and the splitting between them can be estimated as

$$|E_2 - E_1| \leq |E_2 - b(E_2)| + |b(E_2) - b(E_1)| + |b(E_1) - E_1| \leq C_\rho e^{-\rho/\hbar},$$

since  $b(E_1) = b(E_2)$ , for any  $\rho < S_0$  since  $\eta > 0$  is arbitrary.

### Propagation of Quantum Observables

So far we have restricted our analysis to the semiclassical computation of the stable states of the time-independent Schrödinger equation (2). However, there exist some relevant results for what concerns the time evolution operator  $e^{-itH/\hbar}$ , which is the formal solution of the time-dependent Schrödinger equation (1) with Hamiltonian  $H = -\hbar^2 \Delta + V$ . In other words, this result is connected to the time-evolution of quantum observables [8,29]. The main tool is the  $\hbar$ -pseudodifferential calculus we briefly review below.

### Brief Review of $\hbar$ -Pseudodifferential Calculus

Here, we briefly review some basic results of the semiclassical pseudodifferential (also called  $\hbar$ -pseudodifferential) calculus. We refer to the books by Folland [9], Grigis and Sjöstrand [15], Martinez [22] and Robert [28] for a detailed treatment.

In this brief review, in order to avoid some technicalities, we restrict ourselves to the simpler case of bounded potentials; however, some of the following results hold in the general case of unbounded potentials too (see, e.g. [28]). To this end we consider *symbols*  $a(x, y, p) \in S_{3n}(\langle p \rangle^m)$  for some positive  $m$ , that is the function  $a$  defined on  $\mathbb{R}^{3n}$  depends smoothly on  $p$  and for any  $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$  one has

$$|\nabla_p^\alpha a(x, y, p)| = \mathcal{O}(\langle p \rangle^m) \quad (17)$$

uniformly. That is

$$|\nabla_p^\alpha a(x, y, p)| \leq C \langle p \rangle^m = C [1 + |p|]^{m/2}$$

for some positive constant  $C$  independent of  $x, y$  and  $p$ , where  $\nabla_p^\alpha = (\partial_{p_1}^{\alpha_1}, \dots, \partial_{p_n}^{\alpha_n})$  and  $|p| = |p_1| + \dots + |p_n|$ .

We associate to a symbol  $a \in S_{3n}(\langle p \rangle^m)$  the *semiclassical pseudodifferential operator* of degree  $m$  defined for  $u \in C_0^\infty(\mathbb{R}^n)$  as the Fourier integral operator

$$[\text{Op}_\hbar(a)u](x) = \frac{1}{[2\pi\hbar]^n} \cdot \int_{\mathbb{R}^n \times \mathbb{R}^n} e^{i(x-y) \cdot p/\hbar} a(x, y, p) u(y) dy dp.$$

We notice that it is formally self-adjoint on the Hilbert space  $L^2(\mathbb{R}^n)$  when the symbol  $a$  is such that  $a(x, y, p) = \overline{a(y, x, p)}$ .

The above pseudodifferential operator can be extended in a unique way to a linear continuous operator on the space of smooth rapidly decreasing functions  $S(\mathbb{R}^n)$  and its dual space of tempered distributions  $S'(\mathbb{R}^n)$ . Furthermore, when the symbol  $a \in S_{3n}(1)$  then the associate pseudodifferential operator is continuous on  $L^2(\mathbb{R}^n)$  (this result is the so-called Calderón–Villancourt theorem).

It is easy to see that this class of pseudodifferential operators contains the usual differential operators. For instance, in the particular case where the symbol  $a$  has the form

$$a(x, y, p) = \sum_{|\alpha| \leq m} b_\alpha(x) p^\alpha$$

then its associated operator is the differential operator given by

$$\begin{aligned} [\text{Op}_\hbar(a)u](x) &= \left[ \text{Op}_\hbar \left( \sum_{|\alpha| \leq m} b_\alpha(x) p^\alpha \right) u \right](x) \\ &= \sum_{|\alpha| \leq m} b_\alpha(x) (i\hbar \nabla_x)^\alpha u. \end{aligned}$$

The class of pseudodifferential operator is closed with respect to the composition. That is: given two symbols  $a \in S_{3n}(\langle p \rangle^m)$  and  $b \in S_{3n}(\langle p \rangle^{m'})$  then the composition of the associated pseudodifferential operators is still a pseudodifferential operator:

$$\text{Op}_\hbar(a) \circ \text{Op}_\hbar(b) = \text{Op}_\hbar(c)$$

where the symbol  $c \in S_{3n}(\langle p \rangle^{m+m'})$  depends on  $\hbar$  and it is given by the *Moyal product*

$$\begin{aligned} c(x, y, p) &= (a \sharp b)(x, y, p) = \frac{1}{[2\pi\hbar]^n} \\ &\cdot \int_{\mathbb{R}^n \times \mathbb{R}^n} e^{i(x-z) \cdot (\eta-p)/\hbar} a(x, z, \eta) b(z, y, p) dz d\eta \\ &\sim \sum_{|\alpha| \geq 0} \frac{\hbar^{|\alpha|}}{i^{|\alpha|} \alpha!} \nabla_z^\alpha \nabla_\eta^\alpha [a(x, z, \eta) b(z, y, p)] \Big|_{z=x, \eta=p} \end{aligned}$$

as  $\hbar$  goes to zero; we should underline that the above asymptotic formula becomes an exact formula when the symbol  $a$  is polynomial with respect to  $p$  since the sum becomes finite. As a result it follows that for any *elliptic symbol*  $a \in S_{3n}(\langle p \rangle^m)$ , that is such that  $|a(x, y, p)| \geq C \langle p \rangle^m$  for some positive constant  $C$ , then its associate pseudodifferential operator is invertible in the sense that there exists a symbol  $b \in S_{3n}(\langle p \rangle^{-m})$ , depending on  $\hbar$ , where

$$\text{Op}_\hbar(a) \circ \text{Op}_\hbar(b) = 1 + \text{Op}_\hbar(s)$$

and

$$\text{Op}_\hbar(b) \circ \text{Op}_\hbar(a) = 1 + \text{Op}_\hbar(s),$$

with  $r, s = \mathcal{O}(\hbar^\infty)$  in  $S_{3n}(1)$ . The symbol  $b \sim \sum_j \hbar^j b_j$  is iteratively defined where  $b_0 = \frac{1}{a}$ .

Now, we are ready to define the *quantization of a classical observable*. Classical observables are functions  $a(x, p)$  of the position  $x \in \mathbb{R}^n$  and of the momenta  $p \in \mathbb{R}^n$ , if we assume that  $a$  is a smooth function such that

$$\left| \nabla_p^\alpha a(x, p) \right| \leq C \langle p \rangle^m$$

then, for any  $\rho \in [0, 1]$ , it follows that

$$a^\rho(x, y, z) = a[(1-\rho)x + \rho y, p] \in S_{3n}(\langle p \rangle^m).$$

We define the *quantization of the observable*  $a$  as:

$$\text{Op}_\hbar^\rho(a) = \text{Op}_\hbar(a^\rho)$$

where the values  $\rho = 0, \frac{1}{2}, 1$  will play a special role; in particular, for  $\rho = 0$  we have the *standard* (also called *left*) quantization, for  $\rho = 1$  we have the *right* quantization and for  $\rho = \frac{1}{2}$  we have the *Weyl* quantization

$$\text{Op}_\hbar^W(a) := \text{Op}_\hbar^{1/2}(a) = \text{Op}_\hbar(a^{1/2}).$$

We emphasize that the Weyl quantization is particularly important in quantum mechanics because when the classical observable  $a$  is a real valued function then the associate pseudodifferential operator  $\text{Op}_\hbar^W(a)$  is formally self-adjoint on  $L^2(\mathbb{R}^n)$ .

We close this review by recalling the following important result which connects the commutator  $[A, B] = AB - BA$  of the pseudodifferential operators  $A = \text{Op}_\hbar^\rho(a)$  and  $B = \text{Op}_\hbar^\rho(b)$ , where  $a, b$  are two classical observables, and the *Poisson bracket* of  $a$  and  $b$ : let  $c$  be the unique symbol such that  $\text{Op}_\hbar^\rho(c) = [A, B]$ , then

$$c = -i\hbar \{a, b\} + \mathcal{O}(\hbar^2).$$

### Egorov Theorem

Now, we are ready to compare the classical evolution of a classical observable with the quantum evolution of its quantum counterpart. With more details let  $h(x, p)$  be a classical Hamiltonian where  $x \in \mathbb{R}^n$  denotes the spatial variable and  $p \in \mathbb{R}^n$  denotes the momentum. Let

$$\phi^t: \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$$

be the (classical) Hamiltonian flux. Thus, for any classical observable function  $b = b(x, p) \in C^\infty(\mathbb{R}^{2n}; \mathbb{R})$  let

$$b^t(x, p) = (b \circ \phi^t)(x, p) = b[\phi^t(x, p)]$$

be the classical evolution of this observable.

It is well known that we can associate to any real-valued classical observable  $b(x, p)$  a symmetric  $\hbar$ -pseudodifferential linear operator denoted by  $\text{Op}_\hbar^W(b)$  by means of the semiclassical (Weyl) quantization rule formally defined as

$$\begin{aligned} [\text{Op}_\hbar^W(b)u](x) &:= \frac{1}{[2\pi\hbar]^n} \\ &\cdot \int_{\mathbb{R}^n \times \mathbb{R}^n} e^{i(x-y) \cdot p/\hbar} b\left(\frac{x+y}{2}, p\right) u(y) dy dp \end{aligned}$$

where  $(x-y) \cdot p = \sum_{i=1}^n (x_i - y_i) p_i$ . In order to properly define this integral operator on the Hilbert space  $L^2(\mathbb{R}^n)$  we require that estimate (17) holds; actually, to the present purposes it is sufficient to assume the following weaker condition on the observable  $b$ :

$$\left| \nabla_{x,p}^\alpha b(x, p) \right| \leq C [1 + |x|^2 + |p|^2]^{m/2} \quad (18)$$

for some  $m \geq 0$  and any  $\alpha \in \mathbb{N}^{2n}$ .

For instance, if  $h_0$  is the Hamiltonian associated to a harmonic oscillator, then it is a quadratic function with respect to both position and momentum variables

$$h_0(x, p) = \sum_{j=1}^n [p_j^2 + \omega_j^2 x_j^2] \quad (19)$$

and the associated Hamiltonian operator (let  $n = 1$  for the sake of simplicity) takes the usual form

$$\begin{aligned} [H_0 u](x) &= [\text{Op}_h^W(h_0)u](x) \\ &= \frac{1}{2\pi\hbar} \int_{\mathbb{R} \times \mathbb{R}} e^{i(x-y)p/\hbar} \left[ p^2 + \omega^2 \left( \frac{x+y}{2} \right)^2 \right] \\ &\quad \cdot u(y) dy dp \\ &= -\hbar^2 \frac{d^2}{dx^2} + \omega^2 x^2 \end{aligned}$$

by integrating by parts twice and since  $1/(2\pi\hbar) \int_{\mathbb{R}} e^{i(x-y)p/\hbar} dp = \delta(x-y)$ .

In such a way it is possible to associate a classical observable  $b$  to a quantum operator  $B$ . If we denote by  $B = \text{Op}_h^W(b)$  and  $H = \text{Op}_h^W(h)$  the operators associated to the classical observable  $b$  and to the Hamiltonian  $h$  then the time quantum evolution  $B^t = e^{itH/\hbar} B e^{-itH/\hbar}$  of the observable  $B$  solves the Heisenberg equation

$$\frac{dB^t}{dt} = \frac{i}{\hbar} [H, B^t]$$

and it is, in some sense, related with the classical evolution  $b^t$  as we are going to explain.

In the very particular case where the Hamiltonian  $h_0$  is given by the harmonic oscillator (19) then we have that the quantum evolution  $B^t$  of the observable  $B$  is a  $\hbar$ -pseudodifferential operator with symbol  $b^t$  given by means of the classical flux Hamiltonian; that is

$$e^{itH_0/\hbar} \text{Op}_h^W(b) e^{-itH_0/\hbar} = \text{Op}_h^W(b \circ \phi^t)$$

where  $\phi^t$  is the Hamiltonian flux generated by the Hamiltonian (19).

In other words, *for the harmonic oscillator Hamiltonian then quantum evolution and Weyl quantization commute*. This property is specific for this problem and it comes from the fact that the flux Hamiltonian is a linear function with respect to  $(x, p)$ . This fact is not true in a general case. However, it is possible to see that a generalization of such a result holds when  $H_0$  is replaced by any symmetric  $\hbar$ -pseudodifferential operator  $H$ ; this result is the so-called *Egorov theorem* (see [8] for the original statement, see also [29] for a detailed review). In particular, under some assumptions, it is possible to prove that the zeroth order remainder term

$$R(t) := e^{itH/\hbar} \text{Op}_h^W(b) e^{-itH/\hbar} - \text{Op}_h^W(b \circ \phi^t)$$

is a bounded operator such that  $\|R(t)\| = \mathcal{O}(\hbar)$  in the sense that the norm of the operator  $R(t)$  is bounded

$$\|R(t)\| \leq C\hbar$$

for some  $C > 0$  and for any  $t \geq 0$ .

Furthermore, it is possible to extend this asymptotic result to any order  $N \geq 1$  for a suitable choice of  $p_{n,t}$ , that is the classical and quantum evolution coincides up to a term of order  $\hbar^{N+1}$  in the semiclassical limit:

$$\begin{aligned} e^{itH/\hbar} \text{Op}_h^W(b) e^{-itH/\hbar} - \text{Op}_h^W(b \circ \phi^t) \\ - \sum_{n=1}^N \hbar^n \text{Op}_h^W(p_{n,t}) = \mathcal{O}(\hbar^{N+1}) \end{aligned}$$

in the norm sense.

### Future Directions

Semiclassical methods are a field of research where theoretical results are rapidly evolving. Just to name some active research topics:  $\hbar$ -pseudodifferential operators, Weyl functional calculus, frequency sets, semiclassical localization of eigenfunctions, semiclassical resonant states, Born–Oppenheimer approximation, stability of matter and Scott conjecture, semiclassical Lieb–Thirring inequality, Peierls substitution rule, etc. Furthermore, semiclassical methods have been also successfully applied in different contests such as superfluidity and statistical mechanics.

Looking forward, we see new emerging research fields in the area of semiclassical methods: *numerical WKB interpolation techniques* and *semiclassical nonlinear Schrödinger equations*.

Indeed, the recent researches in the area of nanosciences and nanotechnologies have opened up new fields where models for semiconductor devices of increasingly small size and electric charge transport along nanotubes cannot be fully understood without considering their quantum nature. Since the oscillating behavior of the solutions of the Schrödinger equation induces serious difficulties for standard numerical simulations, then new numerical approaches based on WKB interpolation are required [1,4,26].

Although the nonlinear Schrödinger equation has been an argument of theoretical research since the 1970s, only in the last few years, with the successful experiments on Bose–Einstein condensate states, has an increasing interest been shown. When we add a nonlinear term to the time-dependent Schrödinger equation (1) then the dynamics of the model drastically changes and new peculiar features, such as the *blow-up effect* and *stability of*

*stationary states*, appear. Semiclassical arguments applied to nonlinear Schrödinger equations justify their reduction to finite dimensional dynamical systems and thus it is possible to obtain an approximate solution, at least for nonlinear time-dependent Schrödinger equations in small dimensional spaces, typically for  $n = 1$  and  $n = 2$ . The extension of this technique to the case  $n > 2$  and the validity of such an approximation for large times is still an open problem [14,27,30].

## Bibliography

- Ben Abdallah N, Pinaud O (2006) Multiscale simulation of transport in an open quantum system: Resonances and WKB interpolation. *J Comp Phys* 213:288–310
- Berezin FA, Shubin MA (1991) The Schrödinger equation. Kluwer, Dordrecht
- Berry MV, Mount KE (1972) Semiclassical approximation in wave mechanics. *Rep Prog Phys* 35:315–397
- Bonnaillie-Noël V, Nier F, Patel Y (2006) Computing the steady states for an asymptotic model of quantum transport in resonant heterostructures. *J Comp Phys* 219:644–670
- Claviere P, Jona Lasinio G (1986) Instability of tunneling and the concept of molecular structure in quantum mechanics: The case of pyramidal molecules and the enantiomer problem. *Phys Rev A* 33:2245–2253
- Dimassi M, Sjöstrand J (1999) Spectral asymptotics in the semiclassical limit. In: London Math Soc Lecture Note Series 268. Cambridge University Press, Cambridge
- Dingle RB (1973) Asymptotic expansion: Their derivation and interpretation. Academic, London
- Egorov YV (1971) Canonical transformation of pseudo-differential operators. *Trans Moscow Math Soc* 24:1–24
- Folland G (1988) Harmonic analysis in phase space. Princeton University Press, Princeton
- Fröman N, Fröman PO (1965) JWKB approximation. North Holland, Amsterdam
- Fröman N, Fröman PO (2002) Physical problems solved by the phase integral methods. Cambridge University Press, Cambridge
- Graffi S, Grecchi V, Jona-Lasinio G (1984) Tunneling instability via perturbation theory. *J Phys A: Math Gen* 17:2935–2944
- Grecchi V, Martinez A, Sacchetti A (1996) Splitting instability: The unstable double wells. *J Phys A: Math Gen* 29:4561–4587
- Grecchi V, Martinez A, Sacchetti A (2002) Destruction of the beating effect for a non-linear Schrödinger equation. *Comm Math Phys* 227:191–209
- Grigis B, Sjöstrand J (1994) Microlocal analysis for differential operators. An introduction. In: London Math. Soc. Lecture Note Series 196. Cambridge University Press, Cambridge
- Harrell EM (1980) Double wells. *Commun Math Phys* 75:239–261
- Helfffer B (1988) Semi-classical analysis for the Schrödinger operator and applications. Lecture Notes in Mathematics 1336. Springer, Berlin
- Helfffer B, Sjöstrand J (1984) Multiple wells in the semiclassical limit I. *Comm Part Diff Eq* 9:337–408
- Helfffer B, Sjöstrand J (1986) Resonances en limite semi-classique. *Mém Soc Math France (N.S.)* 24–25
- Hislop P, Sigal IM (1996) Introduction to spectral theory. In: Appl Math Sci, vol. 113. Springer, New York
- Landau LD, Lifshitz EM (1959) Quantum mechanics. Course of theoretical physics. Pergamon, Oxford
- Martinez A (2002) An introduction to semiclassical and microlocal analysis. Springer, New York
- McHugh JAM (1971) An historical survey of ordinary linear differential equations with a large parameter and turning points. *Arch Hist Exact Sci* 7:277–324
- Merzbacher E (1970) Quantum mechanics, 2nd edn. Wiley, New York
- Olver FWJ (1974) Asymptotics and Special Functions. Academic, New York
- Presilla C, Sjöstrand J (1996) Transport properties in resonant tunneling heterostructures. *J Math Phys* 37:4816–4844
- Raghavan S, Smerzi A, Fantoni S, Shenoy SR (1999) Coherent oscillations between two weakly coupled Bose–Einstein condensates: Josephson effects,  $\pi$  oscillations, and macroscopic quantum self-trapping. *Phys Rev A* 59:620–633
- Robert D (1987) Autour de l'Approximation Semiclassique. Birkhäuser, Basel
- Robert D (1988) Semi-classical approximation in quantum mechanics. A survey of old and recent mathematical results. *Helv Phys Acta* 71:44–116
- Sacchetti A (2005) Nonlinear double well Schrödinger equations in the semiclassical limit. *J Stat Phys* 119:1347–1382
- Simon B (1983) Semiclassical limit of low lying Eigenvalues I: Non degenerate minima. *Ann H Poincaré* 38:295–307
- Simon B (1985) Semiclassical limit of low lying Eigenvalues IV: The flea of the elephant. *J Funct Anal* 63:123–136
- Voros A (1982) Spectre de l'Équation de Schrödinger et Méthode BKW. Publications Mathématiques d'Orsay 81.09
- Wilkinson M, Hannay JH (1987) Multidimensional tunneling between excited states. *Phys D: Nonlin Phenom* 27:201–212

## Perturbative Expansions, Convergence of

SEBASTIAN WALCHER

Lehrstuhl A für Mathematik, RWTH Aachen, Aachen, Germany

## Article Outline

Glossary

Definition of the Subject

Introduction

Poincaré–Dulac Normal Forms

Convergence and Convergence Problems

Lie Algebra Arguments

NFIM and Sets of Analyticity

Hamiltonian Systems

Future Directions

Bibliography

## Glossary

**Resonant eigenvalues** Let  $B$  be a linear endomorphism of  $\mathbb{C}^n$ , with eigenvalues  $\lambda_1, \dots, \lambda_n$  (counted according to multiplicity). One calls these eigenvalues *resonant* if there are integers  $d_j \geq 0$ ,  $\sum d_j \geq 2$ , and some  $k \in \{1, \dots, n\}$  such that

$$d_1 \lambda_1 + \dots + d_n \lambda_n - \lambda_k = 0.$$

If  $B$  is represented by a matrix in Jordan canonical form then the associated vector monomial  $x_1^{d_1} \dots x_n^{d_n} e_k$  will be called a *resonant monomial*. (Here  $e_1, \dots, e_n$  denote the standard basis, and the  $x_i$  are the corresponding coordinates.)

**Poincaré–Dulac normal form** Let  $f = B + \dots$  be a formal or analytic vector field about 0, and let  $B_s$  be the semisimple part of  $B$ . Then one says that  $f$  is in Poincaré–Dulac normal form (PDNF) if  $[B_s, f] = 0$ . An equivalent characterization, if  $B$  is in Jordan form, is to say that only resonant monomials occur in the series expansion.

**Normalizing transformations and convergence** A relatively straightforward argument shows that any formal vector field  $f = B + \dots$  can be transformed to a formal vector field in PDNF via a formal power series transformation. But for analytic vector fields, the existence of a convergent transformation is not assured. There are two obstacles to convergence: First, the possible existence of small denominators (roughly, this means that the eigenvalues satisfy “near-resonance conditions”); and second, “algebraic” obstructions due to the particular form of the normalized vector field.

**Lie algebras of vector fields** The vector space of analytic vector fields on an open subset  $U$  of  $\mathbb{C}^n$ , with the bracket  $[p, q]$  defined by

$$[p, q](x) := Dq(x)p(x) - Dp(x)q(x)$$

becomes a Lie algebra, as is well known. Mutatis mutandis, this also holds for local analytic and for formal vector fields. As noted previously, PDNF is most naturally defined via this Lie bracket. Moreover, the “algebraic” obstructions to convergence are most appropriately discussed within the Lie algebra framework.

**Normal form on invariant manifolds** While there may not exist a convergent transformation to PDNF for a given vector field  $f$ , one may have a convergent transformation to a “partially normalized” vector field, which admits a certain invariant manifold and is in PDNF when restricted to this manifold. This observation is of some practical importance.

## Definition of the Subject

It seems appropriate to first clarify what types of perturbative expansions are to be considered here. There exist various types of such expansions in various settings (a very readable introduction is given in Verhulst’s monograph [41]), but for many of these settings the question of convergence is not appropriate or irrelevant. Therefore we restrict attention to the scenario outlined, for instance, in the introductory chapter of the monograph [13] by Cicogna and Gaeta, which means consideration of normal forms and normalizing transformation for local analytic vector fields.

Normal forms are among the most important tools for the local analysis and classification of vector fields and maps near a stationary point. (See ► [Normal Forms in Perturbation Theory](#).) Convergence problems arise here, and they turn out to be surprisingly complex. While the first contributions date back more than a century, some very strong and very deep results are just a few years old, and this remains an active area of research. Clearly convergence questions are relevant for the analytic classification of local vector fields, but they are also of practical relevance in applications, e. g. for stability questions and for the existence of particular types of solutions.

## Introduction

The theory of normal forms was initiated by Poincaré [35], and later extended by Dulac [17], and by Birkhoff [5] to Hamiltonian vector fields. There exist various types of normal forms, depending on the specific problem one wants to address. Bruno (see [6,7]) in the 1960s and 1970s performed a comprehensive and deep investigation of Poincaré–Dulac normal forms, which are defined with respect to the semisimple part of the linearization. Such normal forms are very important in applications, and moreover they have certain built-in symmetries, which allows a well-defined reduction procedure.

We will first give a quick review of normalization procedures and normal forms, and then discuss convergence problems (which mostly refers to convergence or divergence of normalizing transformations). We will present fundamental convergence and divergence results due to Poincaré, Siegel, Bruno and others. We then proceed to discuss the relevance of certain Lie algebras of analytic vector fields for these matters, including results by Cicogna and the author of this article on the influence of symmetries, and the far-reaching generalization of Bruno’s theorems (among others) due to Stolovitch. Variants of normal forms which guarantee convergence on certain subsets (due to Bibikov and Bruno) are then discussed, and ap-



plications are mentioned. Finally, the Hamiltonian setting, which deserves a discussion in its own right, is presented, starting with results of Ito and recently culminating in Zung's convergence theorem. For Hamiltonian systems there is also work due to Perez-Marco on divergence of normal forms.

Within the space limitations of this contribution, and in view of some very intricate and space-consuming technical questions and conditions, the author tried to find an approach that, for some problems, should provide some insight into a result, the arguments in its proof or its relevance, without exhibiting all the technicalities, or without giving the most general statement. The author hopes to have been somewhat successful in this, and apologizes to the creators of the original theorems for presenting just "light" versions.

### Poincaré–Dulac Normal Forms

We will start with a coordinate-free approach to normal form theory. Our objects are local ordinary differential equations (over  $\mathbb{K} = \mathbb{R}$  or  $\mathbb{C}$ )

$$\dot{x} = F(x), \quad F(0) = 0$$

with  $F$  analytic, thus we have a convergent series expansion

$$F(x) = Bx + \sum_{j \geq 2} f_j(x) = Bx + f_2(x) + f_3(x) + \dots$$

near  $0 \in \mathbb{K}^n$ .

Here  $B = DF(0)$  is linear, and each  $f_j$  is homogeneous of degree  $j$ . Our objective is to simplify the Taylor expansion of  $F$ . For this purpose, take an analytic "near-identity" map

$$H(x) = x + h_2(x) + \dots$$

Since  $H$  is locally invertible, there is a unique

$$F^*(x) = Bx + \sum_{j \geq 2} f_j^*(x)$$

such that the identity

$$DH(x)F^*(x) = F(H(x)) \quad (\text{R})$$

holds.  $H$  "preserves solutions" in the sense that parametrized solutions of  $\dot{x} = F^*(x)$  are mapped to parametrized solutions of  $\dot{x} = F(x)$  by  $H$ . It is convenient to introduce the following abbreviation:

$$F^* \xrightarrow{H} F \quad \text{if (R) holds.}$$

Given the expansion

$$F(x) = Bx + f_2(x) + \dots + f_{r-1}(x) + f_r(x) + \dots,$$

assume that  $f_2, \dots, f_{r-1}$  are already deemed "satisfactory" (according to some specified criterion). Then the ansatz  $H(x) = x + h_r(x) + \dots$  yields

$$F^*(x) = Bx + f_2(x) + \dots + f_{r-1}(x) + f_r^*(x) + \dots$$

(with terms of degree  $< r$  unchanged), and at degree  $r$  one obtains the so-called *homological equation*:

$$[B, h_r] = f_r - f_r^*$$

(Here  $[p, q](x) = Dq(x)p(x) - Dp(x)q(x)$  denotes the usual Lie bracket of vector fields).

The space  $\mathcal{P}_r$  of homogeneous vector polynomials of degree  $r$  is finite dimensional, and  $\text{ad} B = [B, \cdot]$  sends  $\mathcal{P}_r$  to  $\mathcal{P}_r$ . Thus the homological equation poses a linear algebra problem on a finite dimensional vector space: Given  $B$  and  $f_r$ , determine  $f_r^*$  so that the equation can be solved and let  $h_r$  be a solution. How can  $f_r^*$  be chosen? Let  $\mathcal{W}$  be any subspace of  $\mathcal{P}_r$  such that

$$\text{image}(\text{ad} B) + \mathcal{W} = \mathcal{P}_r.$$

Then one may choose  $f_r^* \in \mathcal{W}$ . If the sum is direct then  $f_r^* \in \mathcal{W}$  is uniquely determined by  $f_r$ .

Generally, the type of normal form is specified by the choice of a subspace  $\mathcal{W}_r$  for each degree  $r$  such that  $\text{image}(\text{ad} B) + \mathcal{W}_r = \mathcal{P}_r$ . The Poincaré–Dulac choice is as follows: Given the decomposition

$$B = B_s + B_n$$

into semisimple and nilpotent part, then  $\text{ad} B = \text{ad} B_s + \text{ad} B_n$  is known to be the corresponding decomposition on  $\mathcal{P}_r$ . Choose  $\mathcal{W}_r = \text{Ker}(\text{ad} B_s)$ . By linear algebra

$$\mathcal{W}_r + \text{image}(\text{ad} B) = \mathcal{P}_r;$$

and the sum is direct if  $B$  is semisimple. In any case we have  $[B_s, f_r^*] = 0$ .

**Definition 1** The vector field  $F^*$  is in *Poincaré–Dulac normal form* if  $[B_s, f_j^*] = 0$  for all  $j$ ; equivalently if  $[B_s, F^*] = 0$ .

If one sets aside convergence questions for the moment, an immediate consequence of the considerations above is:

**Proposition 1** For any  $F = B + \dots$  there are formal power series  $H(x) = x + \dots$ ,  $F^*(x) = Bx + \dots$  such that  $F^* \xrightarrow{H} F$  and  $F^*$  is in Poincaré–Dulac normal form.

One should note that the homological equation is of relevance beyond the setting of Poincaré–Dulac, and forms the foundation for various other (or more refined) types of normal form. See also the entry in this section of the Encyclopedia by T. Gramchev on normal forms with respect to a nilpotent linear part.

Now let us turn to the standard approach via suitable coordinates. Given  $F$  be as above, let  $\lambda_1, \dots, \lambda_n$  be the eigenvalues of  $B$ . Complexify, if necessary, and assume that  $B$  is in Jordan canonical form with respect to the given coordinates  $x_1, \dots, x_n$  (with corresponding basis  $e_1, \dots, e_n$  of  $\mathbb{K}^n$ ). In particular,

$$B_s = \text{diag}(\lambda_1, \dots, \lambda_n).$$

In the following we will use  $x_i$  and  $e_i$  to denote eigencoordinates, resp. eigenbasis elements, and reserve  $\lambda_1, \dots, \lambda_n$  for the eigenvalues of  $B$ , without explicitly saying so in every instance.

**Lemma 1** The “vector monomial”  $p(x) = x_1^{m_1} \dots x_n^{m_n} e_j$  satisfies

$$[B_s, p] = (m_1\lambda_1 + \dots + m_n\lambda_n - \lambda_j) \cdot p$$

Thus, the vector monomials form an eigenbasis of  $\text{ad } B_s$  on the space  $\mathcal{P}_r$ , with eigenvalues  $m_1\lambda_1 + \dots + m_n\lambda_n - \lambda_j$  ( $m_j$  nonnegative integers,  $\sum m_j = r$ ).

The eigenvalues play a crucial role both for the classification of formal normal forms and for convergence issues. The following distinction is pertinent here:

**Definition 2** One calls  $(\lambda_1, \dots, \lambda_n)$  *resonant* if there are integers  $d_j \geq 0$ ,  $\sum d_j \geq 2$ , and some  $k \in \{1, \dots, n\}$  such that

$$d_1\lambda_1 + \dots + d_n\lambda_n - \lambda_k = 0.$$

In this case, the vector monomial  $x_1^{d_1} \dots x_n^{d_n} e_k$  is also called a *resonant monomial*. One calls  $(\lambda_1, \dots, \lambda_n)$  *non-resonant* otherwise.

Given eigencoordinates, one can characterize a Poincaré–Dulac normal form by the property that only resonant monomials occur in the series expansion. Moreover, evaluation of the homological equation shows that the nonzero eigenvalues of  $\text{ad } B$  will occur as denominators in its solution. To iterate the normalization, one may proceed degree by degree with a series of transformations of the form  $\exp(h_r)$ , using the solution of the homological equation; see [44]. One may also choose a different iterative approach to compute a normalizing transformation, such as the important “distinguished

transformation” of Bruno [6]. In any case this will yield coefficients whose denominators are products of nonzero terms  $m_1\lambda_1 + \dots + m_n\lambda_n - \lambda_j$ . This is the source of convergence problems caused by *small denominators*:

There are formal series  $H$  and  $F^*$  such that  $F^* \xrightarrow{H} F$  and  $F^*$  is in Poincaré–Dulac normal form, but does there exist a convergent  $H$ ? (Here “convergent” means: convergent in some neighborhood of 0.)

To answer this question, it is necessary to specify a particular type of transformation: Transformations to Poincaré–Dulac normal form are not necessarily unique, even if one stipulates a near-identity transformation  $H(x) = x + \dots$ . Non-uniqueness occurs whenever the eigenvalues of  $B$  are resonant, because then some homological equation will not have a unique solution: In eigencoordinates of  $B_s$ , the series of the transformation is fixed only up to resonant monomial terms. In particular, if  $F$  itself is in normal form then any formal power series  $H(x) = x + \dots$  that contains only resonant monomials will provide a normal form  $F^*$ , and a suitable nontrivial choice of  $H$  will even force  $F^* = F$ . Thus there may be divergent transformations sending an equation in normal form to itself.

## Convergence and Convergence Problems

Let us note at the start that for given analytic  $F$  there may not exist any convergent transformation to normal form; thus the convergence question has no simple answer.

An early positive convergence result is due to Poincaré [35], with a later improvement due to Dulac [17]:

**Theorem 1 (Poincaré–Dulac)** If  $\lambda_1, \dots, \lambda_n$  lie in an open half-plane in  $\mathbb{C}$  which does not contain 0 then there exists a convergent transformation.

For example, one may think of the open left half-plane. The proof is relatively straightforward, employing natural majorants. (Due to the hypothesis, the  $|\sum_{i=1}^n m_i\lambda_i - \lambda_j|$  are unbounded for  $\sum m_i \rightarrow \infty$ .) Poincaré’s condition does not preclude the existence of resonant monomials but it ensures that there are at most finitely many of these.

One main technical difficulty in proving convergence was to replace the direct majorant arguments by more efficient, and more sophisticated, tools, so that small denominator problems could be tackled. The following result, due to C.L. Siegel [38], may be seen as the start of the “modern phase” for convergence results. Characteristically, this result goes back to a mathematician who also worked in analytic number theory. Siegel assumed that the eigenvalues satisfy a certain arithmetic condition.

**Condition S:** The eigenvalues are pairwise different and there are constants  $C > 0$ ,  $\nu > 0$  such that for all nonnegative integer tuples  $(m_i)$ ,  $\sum m_i > 1$ , the following inequality holds:

$$\left| \sum_{i=1}^n m_i \lambda_i - \lambda_j \right| \geq C \cdot (m_1 + \dots + m_n)^{-\nu}$$

**Theorem 2 (Siegel)** *If Condition S holds then there is a convergent transformation to normal form.*

*Proof* A very rough sketch of the proof is as follows. One works in eigencoordinates. By scaling one may assume that all coefficients in the expansion of  $F$  are absolutely bounded by some constant  $M \geq 1$ . For the transformation one writes

$$H(x) = x + \sum_{r \geq 2} \left( \sum \alpha_{m_1, \dots, m_n, k} \cdot x_1^{m_1} \dots x_n^{m_n} e_k \right)$$

where the sum inside the bracket extends over all nonnegative integer tuples with  $\sum m_i = r$  and  $1 \leq k \leq n$ . (Since Condition S precludes resonances, the coefficients of the series are uniquely determined.) Now set

$$A_{m_1, \dots, m_n} := \sum_k |\alpha_{m_1, \dots, m_n, k}|.$$

From the homological equations one finds by recursion

$$\left| \sum_i m_i \lambda_i - \lambda_k \right| \cdot |\alpha_{m_1, \dots, m_n, k}| \leq M \cdot \sum A_{d_{1,1}, \dots, d_{1,n}} \dots A_{d_{s,1}, \dots, d_{s,n}}$$

where the summation on the right hand side extends over all tuples  $(d_{i,1}, \dots, d_{i,n})$  that add up to  $(m_1, \dots, m_n)$ . From this one may obtain an estimate for  $A_{m_1, \dots, m_n}$ , and invoking Condition S one eventually arrives at the conclusion that  $\sum A_{m_1, \dots, m_n} x_1^{m_1} \dots x_n^{m_n}$  is majorized by the series of

$$\frac{x_1 + \dots + x_n}{1 - K \cdot (x_1 + \dots + x_n)}, \quad \text{some } K > 0.$$

□

**Example** Let  $\dot{x} = Bx + \dots$  be given in dimension two, and assume that the eigenvalues  $\lambda_1, \lambda_2$  of  $B$  are nonresonant, and are algebraic irrational numbers. (This is the case when the entries of  $B$  are rational but the characteristic polynomial is irreducible over the rationals.) Then  $\lambda_2/\lambda_1$  is algebraic but not rational, and  $(\lambda_1, \lambda_2)$  satisfies Condition S, due to a celebrated number-theoretic result

of Thue, Siegel and Roth. Thus there exists a convergent transformation to normal form.

While Siegel's convergence proof uses majorizing series, the approach is not as straightforward as in the Poincaré setting. Siegel's result is strong in the sense that Condition S is satisfied by Lebesgue – almost all tuples  $(\lambda_1, \dots, \lambda_n) \in \mathbb{C}^n$ . But the condition forces the normal form to be uninteresting: One necessarily has  $F^* = B = B_s$ . In the same paper [38], Siegel also notes that divergence is possible, even for a set of “eigenvalue vectors” that is everywhere dense in  $n$ -space.

In the resonant case there is a second source of obstacles to convergence. An early example for this is due to Horn (about 1890); see [6]:

**Example** The system

$$\begin{aligned} \dot{x}_1 &= x_1^2 \\ \dot{x}_2 &= x_2 - x_1 \end{aligned}$$

(with eigenvalues  $(0, 1)$  for the linear part) admits no convergent transformation to normal form. A detailed proof for this can be found in [14]. The underlying reason is that the ansatz for a transformation – unavoidably – leads to the differential equation

$$x^2 \cdot y' = y - x,$$

(which goes back to Euler) with divergent solution  $\sum_{k \geq 1} (k-1)! x^k$ . There are no small denominators here. The problem actually lies within the normal form, which can be computed as

$$\begin{aligned} \dot{y}_1 &= y_1^2 \\ \dot{y}_2 &= y_2. \end{aligned}$$

The single nonlinear term is sufficient to obstruct convergence.

Pliss [34] showed that Siegel's theorem still holds if there are no such nonlinear obstructions in the normal form:

**Theorem 3 (Pliss)** *Assume that:*

- (i) *The nonzero elements among the  $\sum_{i=1}^n m_i \lambda_i - \lambda_j$  satisfy Condition S.*
- (ii) *Some formal normal form of  $F$  is equal to  $B = B_s$ . Then there exists a convergent transformation to normal form.*

While it seemingly extends Siegel's result only to a rather narrow special setting, Pliss' theorem proved to be quite important for future developments. Pliss uses a different approach to proving convergence, via a generalized Newton method.

Fundamental insights into normal forms, and in particular into convergence and divergence problems were achieved by Bruno, starting in the 1960s; see [6,7]. His results included or surpassed much of the earlier work. Let us take a closer look at Bruno's conditions. As above, let  $B$  be in Jordan form, with eigenvalues  $\lambda_1, \dots, \lambda_n$ . For  $k \geq 1$  set

$$\omega_k := \min \left\{ \left| \sum m_i \lambda_i - \lambda_j \right| \neq 0 : \right. \\ \left. 1 \leq j \leq n, m_i \in \mathbb{Z}_+, \sum m_i < 2^k \right\}$$

Bruno introduced two arithmetic conditions:

*Condition  $\omega$ :*

$$-\sum_{k=1}^{\infty} \frac{\ln \omega_k}{2^k} < \infty$$

*Condition  $\bar{\omega}$ :*

$$\limsup_k -\frac{\ln \omega_k}{2^k} < \infty$$

Condition  $\omega$  can be shown to imply Condition S. Clearly Condition  $\omega$  implies Condition  $\bar{\omega}$ .

Possibly in view of Pliss' theorem, Bruno introduced the following algebraic condition on the normal form:

*Condition A:* Some formal normal form is of the type

$$F^* = \sigma \cdot B_s$$

with  $\sigma$  a scalar formal power series.

Pliss' condition corresponds to Condition A with  $\sigma = 1$ . Moreover, one can show that Condition A (as well as Pliss' condition) is satisfied by every (formal) normal form if it is satisfied by one.

Now let us turn to Bruno's main theorems. The convergence theorem here may be expected in view of the previous results, but the divergence theorem – as well as its proof – conquers new ground. (We do not state the most general version of the divergence theorem.)

**Theorem 4 (Bruno's convergence theorem)** *If Condition  $\omega$  and Condition A are satisfied then a convergent normal form transformation exists.*

**Theorem 5 (Bruno's divergence theorem)** *Assume that  $\lambda_1, \dots, \lambda_n$  do not lie in a complex half-plane with 0 in its boundary (in particular they do not satisfy the Poincaré condition). Moreover assume that an analytic vector field  $F^*$  in normal form does not satisfy a weaker version of Condition A, or that Condition  $\bar{\omega}$  is not satisfied. Then there exists an analytic  $F$  with normal form  $F^*$  such that no transformation of  $F$  to (any) normal form converges.*

As for the “weaker version of Condition A”, see Remark (c) below. Bruno also discusses the scenario when all eigenvalues are contained in some complex half-plane with 0 in its boundary.

*Example* There exists an analytic vector field

$$F(x) = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} + \dots$$

which admits no convergent transformation to normal form (which is just the linear part). The arithmetic conditions on the eigenvalues are satisfied, but the nontrivial Jordan block violates Condition A. For the related problem of normalizing local analytic diffeomorphisms (rather than vector fields), see e.g. DeLatte and Gramchev [16], and Gramchev [21], for similar divergence results.

*Remark (a)* Bruno's divergence theorem cannot be applied directly to prove divergence of all normalizing transformations for a specific given vector field. (Therefore any particular example still has to be worked “by hand”.) Rather, the divergence theorem provides a generic result. Since the arithmetic conditions (both  $\omega$  and  $\bar{\omega}$ ) are very weak, one sees that in absence of the Poincaré condition, the algebraic obstructions from the formal normal form are mostly responsible for divergence.

*Remark (b)* Bruno's divergence theorem starts from the assumption that a convergent normal form exists for a given vector field; thus he deals with convergence or divergence of normalizing transformations. Little seems to be known about analytic vector fields that admit only divergent normal forms.

*Remark (c)* The version of Condition A stated here is taken from Bruno's monograph (see Chapter III, § in [7]). The original paper (p. 140 ff. in [6]). contains somewhat different versions. In the interesting case when the hypothesis of Theorem 5 holds, Bruno's original requirement on the normal form in eigencoordinates is as follows:

$$\dot{x}_j = x_j \left( \lambda_j \sigma + \bar{\lambda}_j \tau \right), \quad 1 \leq j \leq n,$$

with scalar series  $\sigma$  and  $\tau$ . This condition seems to have been stated with special regard to real systems, because otherwise complex conjugation plays no distinguished role. To illustrate this point, consider the complexification of a real system, and multiply this through by some scalar  $\exp(i\theta)$  which is neither real nor purely imaginary. Convergence or divergence of normalizing transformations is not affected by this, but the shape of the condition

above changes considerably. (The appropriate setting for this seems to be the one developed by Stolovitch [39]; see below.) In later publications, notably in [7], Bruno himself mostly used Condition A in the simple form we stated above.

### Lie Algebra Arguments

Bruno's theorems set the standard against which later results are measured. However, they do not directly address this basic question: Given a particular local analytic vector field  $F$ , characterize properties that are necessary (and, ideally, also sufficient) for the existence of a convergent normal form transformation. (Obviously, Condition A is not generally necessary for the existence of a convergent normalizing transformation.) To answer this basic question and related problems, it turns out helpful to consider Lie algebras of analytic vector fields.

A possible approach is based on the earlier observation that normal forms admit symmetries. (See also the entry on Symmetry and Perturbation Theory in this section.) The coordinate-free approach proves to be quite suitable here. First, let us formalize the observation:

**Lemma 2** *If  $F$  admits a convergent normalizing transformation then there exists a nontrivial  $G$  (i. e.,  $G \notin \mathbb{K} \cdot F$ ) such that  $[G, F] = 0$ .*

*Proof* There exist a convergent  $\Psi$  and a convergent  $F^*$  in normal form such that

$$F^* \xrightarrow{\Psi} F$$

Now  $\Psi$  sends  $B_s$  to some analytic  $G = B_s + \dots$ , and  $[B_s, F^*] = 0$  implies  $[G, F] = 0$ . If  $F^* \neq B_s$ , we are done. If  $F^* = B_s$  then take some linear map  $C \notin \mathbb{K} \cdot B_s$  that commutes with  $B_s$ , and define  $G$  by  $C \xrightarrow{\Psi} G$ .  $\square$

In other words, if there is a convergent normalizing transformation then there exists a nontrivial infinitesimal symmetry at the stationary point. One can try to turn this necessary condition around and thus obtain sufficient convergence criteria. This has been done, with some success, since the early 1990s. Among the relevant contributions are those by Markhashov [29], Bruno and Walcher [9], Cicogna [11,12], Bambusi et al. [3], and, from a different starting point, Stolovitch [39] (see below and see also Gramchev and Yoshino [24] for maps). The survey paper [14] by Cicogna and Walcher collects the development up to 2001.

Here we will present a few results to give the reader an impression of the arguments employed.

The objects to deal with are  $C_{\text{for}}(F)$  and  $C_{\text{an}}(F)$ , i. e., the formal, respectively analytic, centralizer of  $F$ , which by definition consist of all formal, respectively analytic, vector fields  $H$  such that  $[H, F] = 0$ . The first result is due to Cicogna [11,12] and Walcher [45]. Note that Pliss' theorem (and thus Condition A) plays a crucial role in the proof.

**Theorem 6** *Given the analytic vector field  $F$  with formal normal form  $\widehat{F}$ , assume that*

$$\dim C_{\text{for}}(\widehat{F}) = k < \infty.$$

*If the eigenvalues of  $B$  satisfy Condition  $\omega$  and  $\dim C_{\text{an}}(F) = k$  then there exists a convergent transformation to normal form.*

*Proof* We have  $\dim C_{\text{for}}(F) = \dim C_{\text{for}}(\widehat{F})$ , so  $\dim C_{\text{an}}(F) = k$  implies  $C_{\text{an}}(F) = C_{\text{for}}(F)$ . Given a formal transformation  $\Psi$  with  $\widehat{F} \xrightarrow{\Psi} F$ , there exists an analytic  $H$  such that  $B \xrightarrow{\Psi} H$  and  $[H, F] = 0$ , since  $B \in C_{\text{for}}(\widehat{F})$ . Note that  $H = B + \dots$ , so  $B$  is a normal form of  $H$ . Due to Pliss' theorem, there is a convergent  $\Phi$  with  $B \xrightarrow{\Phi} H$ . Now  $\widetilde{F} \xrightarrow{\Phi} F$  for some  $\widetilde{F}$ , and  $[B, \widetilde{F}] = 0$ , so  $\widetilde{F}$  is in normal form.  $\square$

The dimension of the formal centralizer is computable in many cases. The requirement on Condition  $\omega$  can be relaxed; see [11,45] and [14]. The next result is due to Markhashov [29] for the non-resonant case, and to Bruno and Walcher [9] in the resonant case. One may base a proof on the observation that  $\dim C_{\text{for}}(F)$  is infinite only if Condition A holds, and use Theorem 6 otherwise.

**Theorem 7** *In dimension  $n = 2$ , there is a convergent transformation of  $F$  to normal form if and only if  $F$  admits a nontrivial commuting vector field in 0.*

In dimension two, there are other, very precise, characterizations of resonant vector fields (for  $\lambda_2/\lambda_1$  a negative rational number) admitting a convergent normalization, which can be drawn from the work of Martinet and Ramis [30]. Beyond this, building on work of Ecalle [18,19] and Voronin [42], Martinet and Ramis succeeded in giving an analytical classification of germs of such vector fields, and those admitting a convergent normalizing transformation can be characterized by the vanishing of infinitely many analytical invariants. In this sense, the convergence problem was settled earlier, at least for the interesting cases. But Theorem 7 approaches the question from a different perspective, gives an algebraic characterization and provides structural insight that is not directly available from [30].

Lie algebra arguments also play a fundamental role, from a different perspective, in the work of Stolovitch [39].



We will here give a simplified and incomplete account of his important results; a full presentation would take up much more space. To motivate Stolovitch's approach, note that in many cases there are natural decompositions  $B = B_1 + \dots + B_\ell$ , with all  $[B_i, B_j] = 0$  which come from eigenvalues splitting up into groups of pairwise commensurable ones, with pairwise incommensurability between the groups. The resonance conditions  $\sum m_j \lambda_j - \lambda_k = 0$  then split up into conditions involving only the groups. We illustrate this with a simple example for such a decomposition:

$$\text{diag}(1, -1, \sqrt{2}, -\sqrt{2}) = \\ \text{diag}(1, -1, 0, 0) + \text{diag}(0, 0, \sqrt{2}, -\sqrt{2}).$$

The setting considered by Stolovitch is as follows: Given (complex) analytic vector fields

$$F_1 = B_1 + \dots, \quad \dots, \quad F_\ell = B_\ell + \dots$$

that commute pairwise, thus all  $[F_i, F_j] = 0$ , ask about simultaneous analytic normalization of these vector fields. There are sensible notions of normal form of a vector field with respect to a linear Lie algebra, and in particular there is a natural extension of Poincaré–Dulac for abelian Lie algebras of diagonal matrices; see [39], Sect. “[Convergence and Convergence Problems](#)”: Each semisimple linear part  $B_{i,s}$  commutes with each  $F_j$  in such a normal form. To avoid trivial scenarios, Stolovitch requires the semisimple parts  $B_{i,s}$  to be linearly independent. To formulate diophantine conditions extending Bruno's Condition  $\omega$ , one may proceed as follows: Let  $\lambda_{i,1}, \dots, \lambda_{i,n}$  denote the eigenvalues of  $B_i$ . For nonnegative integers  $m_1, \dots, m_n$ , and for  $1 \leq d \leq n$  set

$$\gamma_{m_1, \dots, m_n, d} := \sum_i \left| \sum_j m_j \lambda_{i,j} - \lambda_d \right|$$

and (for instance; Stolovitch gives a more general formulation)

$$\omega_k(B_1, \dots, B_\ell) := \inf \left\{ \gamma_{m_1, \dots, m_n, d} \neq 0 : \right. \\ \left. 1 \leq d \leq n, 2 \leq \sum m_j \leq 2^k \right\}.$$

This leads to an appropriate diophantine condition which extends Bruno's Condition  $\omega$ .

Condition  $\omega^\#$ :

$$-\sum_{k=1}^{\infty} \frac{\omega_k(B_1, \dots, B_\ell)}{2^k} < \infty$$

We remark briefly that there appear to be some issues of well-definedness here. For instance, the choice of the  $F_i$ , and hence of the  $B_i$ , is not unique, and one has to verify that the important notions, like Condition  $\omega^\#$ , do not depend on these choices. Stolovitch [39] works in an invariant setting from the start; as a consequence, no such questions arise.

Finally, Stolovitch introduces the notion of *formal complete integrability*. Disregarding technical subtleties (even though these are relevant and of interest), one may informally characterize this property as follows: The system  $F_1, \dots, F_\ell$  is formally completely integrable if it has as many formal integrals as admissible by the semisimple linear parts  $B_{1,s}, \dots, B_{\ell,s}$ . To cast at least some light on this, we note that every formal integral of the system in normal form is also a simultaneous first integral of the  $B_{i,s}$ . See Walcher [43] for the case  $\ell = 1$  and Stolovitch [39], Sect. “[Lie Algebra Arguments](#)” for the general case. This notion of complete integrability is the appropriate extension of Bruno's Condition A.

The following is a simplified representative of the results in [39]. The system  $B_1, \dots, B_\ell$  is said to have *small divisors* if 0 is a limit point of the  $\gamma_{m_1, \dots, m_n, d} \neq 0$ .

**Theorem 8 (Stolovitch)** *In the presence of small divisors, if Condition  $\omega^\#$  holds then every formally completely integrable system  $F_1, \dots, F_\ell$  admits a convergent transformation to normal form.*

The principal value of Stolovitch's results is that they open up a unified approach to a number of applications. One almost immediate application is a recovery of Bruno's convergence theorem. (See also Remark (c) following Theorem 5.) We give two more applications. For the first one, compare also Bambusi et al. [3].

**Corollary 1 (Linearization)** *In the presence of small divisors, assuming that Condition  $\omega^\#$  holds, every formally linearizable system  $F_1, \dots, F_\ell$  admits a convergent linearization.*

The second application is a short and clear proof of a theorem due to Vey [40]:

**Corollary 2 (Volume-preserving vector fields)** *Assume that  $\ell = n - 1$  and that  $F_1, \dots, F_{n-1}$  are commuting volume-preserving vector fields, with diagonal and linearly independent linear parts  $B_i = B_{i,s}$ . Then the  $F_i$  are simultaneously analytically normalizable.*

*Remark (a)* For similar results see Zung [46]. Zung uses a different approach based on a convergence result (a precursor of [47]) for Hamiltonian vector fields; see also the section on Hamiltonian vector fields below. His argument seems somewhat sketchy.

*Remark (b)* There is a nontrivial overlap of Stolovitch's results with the approach to convergence via symmetries outlined above: Some results can be proven by either method, as the references indicate.

We finish this section with a third aspect of Lie algebra arguments in convergence proofs. The following is taken from Walcher [44]:

**Theorem 9** *Let  $\mathcal{L}$  be a finite dimensional Lie algebra of polynomial vector fields which is graded in the sense that it contains all homogeneous parts of each of its elements, and let  $F \in \mathcal{L}$  and  $F(0) = 0$ . Then  $F$  admits a convergent transformation to normal form  $F^*$ , and one may take  $F^* \in \mathcal{L}$ .*

The basic idea for the proof is to take suitable solutions  $h_r$  of the homological equations, which can be chosen in  $\mathcal{L}$ , and transformations  $\exp(h_r)$ ; see [44], Prop. 2.7. Due to finite dimension of  $\mathcal{L}$ , finitely many such transformations suffice. There are several interesting Lie algebras among the finite dimensional graded ones:

*Example (a)* Every projective vector field, as well as every conformal vector field in dimension  $\geq 3$ , admits a convergent transformation to Poincaré–Dulac normal form.

*Example (b)* Matrix Riccati equations: Every matrix differential equation of the form

$$\dot{x} = xax + bx + xc,$$

with  $x, a, b, c$  matrices of appropriate sizes, admits a convergent transformation to normal form.

### NFIM and Sets of Analyticity

As we have seen, convergence problems for normalizing transformations are unsurmountable in many instances. But there are sensible strategies to achieve convergence by relaxing the requirements on the normalized vector field. Moreover, such strategies frequently provide interesting information, e.g. on stability or on the existence of periodic solutions. There are two related, but not equivalent, approaches to be discussed here: Bibikov's *normal form on an invariant manifold* (NFIM), see [4]; and Bruno's *sets of analyticity*, see [7].

We will first discuss Bibikov's work, including a coordinate-free approach proposed in [44].

**Definition 3** Let  $C$  be a semisimple linear map. A vector subspace of  $\mathbb{K}^n$  is called *strongly  $C$ -stable* if it is invariant for every vector field  $F = B + \dots$ , with  $B_s = C$ , in normal form.

Strongly  $C$ -stable spaces are in particular  $C$ -stable. A coordinate-dependent characterization is as follows.

**Lemma 3** *Assume that  $C$  is in diagonal form, and let  $1 < r < n$ . Then the subspace  $U := \mathbb{K}e_1 + \dots + \mathbb{K}e_r$  is strongly  $C$ -stable if and only if*

$$m_1\lambda_1 + \dots + m_r\lambda_r - \lambda_j \neq 0$$

*for all nonnegative integers  $m_i$  with  $\sum m_i > 0$ , and all  $j \in \{r+1, \dots, n\}$ .*

Here, the choice of indices  $1, \dots, r$  is just for the sake of convenience. The proof is simple: The condition ensures that no monomial

$$x_1^{m_1} \dots x_r^{m_r} e_j, \quad j > r$$

will occur in the normal form, and therefore the subspace  $U$ , characterized by  $x_{r+1} = \dots = x_n = 0$ , is invariant for any normal form  $F = C + B_n + \dots$ . Examples include the stable, unstable and center subspaces of  $C$ .

Now one can introduce the notion of NFIM:

**Definition 4** Assume that  $U = \mathbb{K}e_1 + \dots + \mathbb{K}e_r$  is strongly  $C$ -stable. A vector field  $F = B + \dots$  with  $B_s = C$  is said to be in *normal form on the invariant manifold  $U$*  (NFIM on  $U$ ) if  $U$  is invariant for  $F$  and furthermore

$$[B_s|_U, F|_U] = 0.$$

*Example* The two-dimensional vector field

$$F = \begin{pmatrix} x_1^2 - 2x_1x_2 + 3x_1^3 \\ x_2(1 + x_1 + x_2^2) \end{pmatrix}, \quad \text{with } Bx = \begin{pmatrix} 0 \\ x_2 \end{pmatrix},$$

is in NFIM on  $U = \mathbb{K}e_1$ .

Bibikov also introduced a refined version of NFIM, which he calls quasi-normal form (QNF). The above example is not in quasi-normal form; in a QNF the first entry would contain only functions of  $x_1$ .

Because normal forms are, in particular, in NFIM on every strongly  $B_s$ -stable subspace, it is obvious that formal transformations to NFIM exist. But there is more freedom to construct such transformations, which may be utilized to force convergence. To give a flavor of Bibikov's results, we present a weakened version of one of his theorems.

**Theorem 10 (Bibikov)** *Given a vector field  $F = B + \dots$ , assume that the subspace  $U := \mathbb{K}e_1 + \dots + \mathbb{K}e_r$  is strongly  $B_s$ -stable, and moreover that:*

(i) *There is an  $\epsilon > 0$  such that*

$$|m_1\lambda_1 + \dots + m_r\lambda_r - \lambda_j| \geq \epsilon$$

*for all nonnegative integers  $m_i$ ,  $\sum m_i > 0$ , and for all  $j > r$ .*

(ii) *Some formal normal form  $\widehat{F}|_U$  satisfies the Pliss condition on  $U$ .*

*Then there exists a convergent transformation to NFIM on  $U$ .*

**Remark** In Bibikov's original theorems (see [4], Theorem 3.2 and Theorem 10.2), one finds a more general (quite technical) condition instead of (ii). Thus the range of Bibikov's theorems is wider than our statement indicates. Condition (i), or some related condition, cannot be discarded completely: There are examples of strongly  $C$ -stable subspaces which do not correspond to analytic invariant manifolds for certain vector fields. (This is another incarnation of the small denominator problem.)

Applications of Bibikov's theorems include the existence of analytic stable and unstable manifolds; stability of the stationary point in case  $\lambda_1 = 0$  when all other  $\lambda_i$  have negative real parts, and  $\widehat{F}|_U = 0$  on  $U = \mathbb{K}e_1$  ("transcendental case"); and the existence of certain periodic solutions.

Let us now turn to Bruno's method of analytic invariant sets; see Bruno Part I, Ch. III, and Part II in [7]. To motivate the approach, one may use the following observation: For an analytic vector field  $\widehat{F} = B + \dots$  in normal form the set

$$\mathcal{A} = \{z : \widehat{F}(z) \text{ and } B_s z \text{ linearly dependent in } \mathbb{K}^n\}$$

is invariant for  $\widehat{F}$ . (This is a consequence of  $[B_s, \widehat{F}] = 0$ : The set of points for which a vector field and an infinitesimal symmetry "point in the same direction" is invariant.) It is clearly possible to write down analytic functions such that  $\mathcal{A}$  is their common zero set: For instance, take suitable  $2 \times 2$ -determinants. One may now introduce the notion that a vector field  $\widehat{F}$  is *normalized on  $\mathcal{A}$* : In the coordinate version this means that for each entry of the right-hand side the sum of all nonresonant parts lies in the defining ideal of  $\mathcal{A}$ .

If  $F$  is not in normal form but some formal normal form  $\widehat{F}$  satisfies Condition A then – assuming some mild diophantine conditions –  $\mathcal{A}$  is a whole neighborhood of the stationary point, according to Bruno's convergence theorem. Bruno now refines this by investigating whether the generally "formal" set  $\mathcal{A}$  is analytic (or at least certain subsets are), and what can be said about the solutions on such subsets. To be more precise: Given a not necessarily convergent normal form  $\widehat{F}$  of  $F$ , one can still write down formal power series whose "common zero set" defines  $\mathcal{A}$ , and it makes sense to ask what of this can be salvaged for analyticity. The following is a sample of Bruno's results (see Part I, Ch. III, Theorem 2 and Theorem 4 in [7]).

As above, we do not write down the technical conditions completely.

**Theorem 11 (Bruno)** *Let the analytic vector field  $F = B + \dots$  be given.*

- (a) *If the eigenvalues  $\lambda_1, \dots, \lambda_n$  of  $B$  are commensurable (i.e., pairwise linearly dependent over the rationals) then the set  $\mathcal{A}$  is analytic and there is a convergent transformation to a vector field  $\tilde{F}$  that is normalized on  $\mathcal{A}$ .*
- (b) *Generally, there exists a (formal) subset  $\mathcal{B}$  of  $\mathcal{A}$  which is analytic, and for  $\mathcal{B}$  the same conclusion as above holds.*

For applications see Bruno [7], Part II, and the recent papers by Edneral [20], and Bruno and Edneral [8], on existence of periodic solutions for certain equations.

## Hamiltonian Systems

Hamiltonian systems have a special position among differential equations (see e.g. [2] for an overview, and ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#)); in view of their importance it is appropriate to give them particular attention. When discussing normal forms of Hamiltonian systems (where Poincaré–Dulac becomes Birkhoff; see [5]) it is natural to consider canonical transformations only. In view of the correspondence between integrals of  $F$  and Hamiltonian vector fields commuting with  $F$ , it is furthermore natural to consider integrals in the Hamiltonian setting. As for convergence results, we first state two theorems by H. Ito [26,27], from around 1990:

**Theorem 12 (Ito; non-resonant case)** *Let  $F = B + \dots$  be Hamiltonian and let  $(\omega_1, -\omega_1, \dots, \omega_r, -\omega_r)$  be the eigenvalues of  $B$ . Moreover assume that  $\omega_1, \dots, \omega_r$  are non-resonant, thus  $\sum m_j \omega_j = 0$  for integers  $m_1, \dots, m_r$  implies  $m_1 = \dots = m_r = 0$ . If  $F$  possesses  $r$  independent integrals in involution (i.e. with vanishing Poisson brackets) then there exists a convergent canonical transformation of  $f$  to Birkhoff normal form.*

This condition is also necessary in the non-resonant case: If there is a convergent transformation to analytic normal form  $\widehat{F}$  then there are  $r$  linearly independent linear Hamiltonian vector fields that commute with  $\widehat{F}$ , and these, in turn, correspond to  $r$  independent quadratic integrals of  $\widehat{F}$ . For the "single resonance" case one has:

**Theorem 13 (Ito; a simple-resonance case)** *Let  $F = B + \dots$  be Hamiltonian and let  $(\omega_1, -\omega_1, \dots, \omega_r, -\omega_r)$  be the eigenvalues of  $B$ . Moreover assume that there are*

nonzero integers  $n_1, n_2$  such that  $n_1\omega_1 + n_2\omega_2 = 0$ , but there are no further resonances. If  $F$  possesses  $r$  independent integrals in involution then there exists a convergent canonical transformation to Birkhoff normal form.

Again, the condition is also necessary. Ito's proofs are quite long and intricate. Kappeler, Kodama and Nemeš [28] proved a generalization of Theorem 13 for more general single-resonance cases. Moreover, they showed that there is a natural obstacle to further generalizations, since there exist non-integrable (polynomial) Hamiltonian systems in normal form. Thus, a complete integrability condition is not generally necessary for convergence.

Nguyen Tien Zung [47] recently succeeded in a far-reaching generalization of Ito's theorems. Considering the existence of non-integrable normal forms, this seems the best possible result on integrability and convergence.

**Theorem 14 (Zung)** *Any analytically integrable Hamiltonian system near a stationary point admits a convergent transformation to Birkhoff normal form.*

One remarkable feature of Zung's proof is its relative shortness, compared with the proofs by Ito.

Finally, turning to divergent normal forms (rather than normalizing transformations) of Hamiltonian systems, Perez-Marco [33] recently established a theorem about convergence or generic divergence of the normal form in the non-resonant scenario. Although numerical computations indicate the existence of analytic Hamiltonian vector fields which admit only divergent normal forms, there still seems to be no example known. Perez-Marco showed that if there is one example then divergence is generic.

### Future Directions

There are various ways to extend the approaches and results presented above, and clearly there are unresolved questions. In the following, some of these will be listed, respectively, recalled.

The convergence problem for normal forms and normalizing transformations is part of the much bigger problem of analytic classification of germs of vector fields. Except for dimension two (see the references [18,19,30] mentioned above) little seems to be known in the case of nontrivial formal normal forms. Going beyond analyticity, an interesting extension would be towards Gevrey spaces. Some work on this topic exists already; see e. g. [22].

Passing from vector fields to maps, matters turn out to be much more complicated in the case of nontrivial formal

normal forms, and even one-dimensional maps show very rich behavior; see [31,32]. A brief introduction is given in the survey [23] mentioned above.

A complete ("algebraic") understanding of Bruno's Condition A would be desirable; this could also provide an approach to a non-Hamiltonian version of Zung's Theorem 14. It seems well possible that all the necessary ingredients for this endeavor are contained in Stolovitch's work [39].

An extension or refinement of Bibikov's and Bruno's results on the existence of certain invariant sets (in the case of non-convergence) would obviously be interesting. There seems to be a natural guideline here: Check what invariant sets are forced onto a system in PDNF and see which ones can be salvaged. (The existence of a commuting vector field, for instance, has more consequences than those exploited by Bruno in the arguments leading up to Theorem 11.)

Finally, one could turn to more refined versions of normal forms, such as normal forms with respect to a nilpotent linear part (see [15]), and quite general constructions such as presented by Sanders [36,37]. It seems that little attention has been paid to convergence questions for such types of normal forms. For normal forms with respect to a nilpotent linear part, there are obviously no small denominator problems, but algebraic obstructions abound. There exists a precise algebraic characterization for such normal forms, involving the representation theory of  $sl(2)$  (see [15]). Thus there may be some hope for an algebraic characterization of convergently normalizable vector fields.

### Bibliography

1. Arnold VI (1982) Geometrical methods in the theory of ordinary differential equations. Springer, Berlin
2. Arnold VI, Kozlov VV, Neishtadt AI (1993) Mathematical aspects of classical and celestial mechanics. In: Arnold VI (ed) Dynamical Systems III, Encyclop Math Sci, vol 3, 2nd edn. Springer, New York
3. Bambusi D, Cicogna G, Gaeta G, Marmo G (1998) Normal forms, symmetry and linearization of dynamical systems. J Phys A 31:5065–5082
4. Bibikov YN (1979) Local theory of nonlinear analytic ordinary differential equations. Lecture Notes in Math 702. Springer, New York
5. Birkhoff GD (1927) Dynamical systems, vol IX. American Mathematical Society, Colloquium Publications, Providence, RI
6. Bruno AD (1971) Analytical form of differential equations. Trans Mosc Math Soc 25:131–288
7. Bruno AD (1989) Local methods in nonlinear differential equations. Springer, New York
8. Bruno AD, Edneral VF (2006) The normal form and the integrability of systems of ordinary differential equations. (Rus-

- sian) Programmirovaniye 3:22–29; translation in Program Comput Software 32(3):139–144
9. Bruno AD, Walcher S (1994) Symmetries and convergence of normalizing transformations. *J Math Anal Appl* 183:571–576
  10. Chow S-N, Li C, Wang D (1994) Normal forms and bifurcations of planar vector fields. Cambridge Univ Press, Cambridge
  11. Cicogna G (1996) On the convergence of normalizing transformations in the presence of symmetries. *J Math Anal Appl* 199:243–255
  12. Cicogna G (1997) Convergent normal forms of symmetric dynamical systems. *J Phys A* 30:6021–6028
  13. Cicogna G, Gaeta G (1999) Symmetry and perturbation theory. In: *Nonlinear Dynamics, Lecture Notes in Phys* 57. Springer, New York
  14. Cicogna G, Walcher S (2002) Convergence of normal form transformations: The role of symmetries. *Acta Appl Math* 70:95–111
  15. Cushman R, Sanders JA (1990) A survey of invariant theory applied to normal forms of vectorfields with nilpotent linear part, *IMA vol Math Appl* 19. Springer, New York, pp 82–106
  16. DeLatte D, Gramchev T (2002) Biholomorphic maps with linear parts having Jordan blocks: linearization and resonance type phenomena. *Math Phys Electron J* 8(2):27 (electronic)
  17. Dulac H (1912) Solutions d'un système d'équations différentielles dans le voisinage de valeurs singulières. *Bull Soc Math Fr* 40:324–383
  18. Ecalle J (1981) Sur les fonctions résurgentes I. *Publ Math d'Orsay* 81(5):1–247
  19. Ecalle J (1981) Sur les fonctions résurgentes II. *Publ Math d'Orsay* 81(6):248–531
  20. Edneral VF (2005) Looking for periodic solutions of ODE systems by the normal form method. In: *Differential equations with symbolic computation*. Trends Math, Birkhäuser, Basel, pp 173–200
  21. Gramchev T (2002) On the linearization of holomorphic vector fields in the Siegel domain with linear parts having nontrivial Jordan blocks SPT. In: *Symmetry and perturbation theory (Cala Gonone)*. World Sci Publ, River Edge, NJ, pp 106–115
  22. Gramchev T, Tolis E (2006) Solvability of systems of singular partial differential equations in function spaces. *Integral Transforms Spec Funct* 17:231–237
  23. Gramchev T, Walcher S (2005) Normal forms of maps: Formal and algebraic aspects. *Acta Appl Math* 87(1–3):123–146
  24. Gramchev T, Yoshino M (1999) Rapidly convergent iteration method for simultaneous normal forms of commuting maps. *Math Z* 231:745–770
  25. Ioss G, Adelmeyer M (1992) Topics in Bifurcation Theory and Applications. World Scientific, Singapore
  26. Ito H (1989) Convergence of Birkhoff normal forms for integrable systems. *Comment Math Helv* 64:412–461
  27. Ito H (1992) Integrability of Hamiltonian systems and Birkhoff normal forms in the simple resonance case. *Math Ann* 292:411–444
  28. Kappeler T, Kodama Y, Nemethi A (1998) On the Birkhoff normal form of a completely integrable system near a fixed point in resonance. *Ann Scuola Norm Sup Pisa Cl Sci* 26:623–661
  29. Markhashov LM (1974) On the reduction of an analytic system of differential equations to the normal form by an analytic transformation. *J Appl Math Mech* 38:788–790
  30. Martinet J, Ramis J-P (1983) Classification analytique des équations différentielles non linéaires résonnantes du premier ordre. *Ann Sci Ecole Norm Sup* 16:571–621
  31. Perez Marco R (1995) Nonlinearizable holomorphic dynamics having an uncountable number of symmetries. *Invent Math* 119:67–127
  32. Perez-Marco R (1997) Fixed points and circle maps. *Acta Math* 179:243–294
  33. Perez-Marco R (2003) Convergence and generic divergence of the Birkhoff normal form. *Ann Math* 157:557–574
  34. Pliss VA (1965) On the reduction of an analytic system of differential equations to linear form. *Differ Equ* 1:153–161
  35. Poincaré H (1879) Sur les propriétés des fonctions définies par les équations aux différences partielles. These, Paris
  36. Sanders JA (2003) Normal form theory and spectral sequences. *J Differential Equations* 192:536–552
  37. Sanders JA (2005) Normal form in filtered Lie algebra representations. *Acta Appl Math* 87:165–189
  38. Siegel CL (1952) Über die Normalform analytischer Differentialgleichungen in der Nähe einer Gleichgewichtslösung. *Nachr Akad Wiss Göttingen, Math-Phys Kl*, pp 21–30
  39. Stolovitch L (2000) Singular complete integrability. *IHES Publ Math* 91:133–210
  40. Vey J (1979) Algèbres commutatives de champs de vecteurs isochores. *Bull Soc Math France* 107(4):423–432
  41. Verhulst F (2005) Methods and applications of singular perturbations Boundary layers and multiple timescale dynamics. In: *Texts in Applied Mathematics*, vol 50. Springer, New York
  42. Voronin SM (1981) Analytic classification of germs of conformal mappings  $(\mathbb{C}, 0) \rightarrow (\mathbb{C}, 0)$ . *Funct Anal Appl* 15:1–13
  43. Walcher S (1991) On differential equations in normal form. *Math Ann* 291:293–314
  44. Walcher S (1993) On transformations into normal form. *J Math Anal Appl* 180(2):617–632
  45. Walcher S (2000) On convergent normal form transformations in presence of symmetries. *J Math Anal Appl* 244:17–26
  46. Zung NT (2002) Convergence versus integrability in Poincaré–Dulac normal form. *Math Res Lett* 9(2–3):217–228
  47. Zung NT (2005) Convergence versus integrability in Birkhoff normal form. *Ann Math (2)* 161(1):141–156

## Phase Transitions in Cellular Automata

NINO BOCCARA<sup>1,2</sup>

<sup>1</sup> Department of Physics, University of Illinois, Chicago, USA

<sup>2</sup> CE Saclay, Gif-sur-Yvette, France

### Article Outline

Glossary

Definition of the Subject

Introduction

The Domany–Kinzel Cellular Automaton

Car Traffic Models



Epidemic Models  
 Future Directions  
 Bibliography

## Glossary

**Phase transition** In statistical physics, a phase transition is the transformation of a macroscopic system from one phase to another. Phase transitions are divided into two types. First-order phase transitions are characterized by discontinuities of first-order derivatives of the Gibbs free energy, such as the entropy or the volume whereas second-order phase transitions are characterized by power-law behaviors of second-order derivatives of the Gibbs free energy such as the specific heat or the magnetic susceptibility. These singular behaviors occur for specific values of intensive variables such as the temperature or the magnetic field. Similar behaviors have also been discovered in many-component systems.

**Critical behavior** Critical behavior manifests itself in many-component systems and is characteristic of a cooperative behavior of the various components. This notion has been introduced in equilibrium statistical physics for many-body systems that exhibit second-order phase transitions. In the vicinity of the transition temperature  $T_c$ , a singular physical quantity  $Q$  has a power-law behavior, that is, it behaves as  $|T - T_c|^\varepsilon$ , where  $\varepsilon$  is the critical exponent characterizing the critical behavior of  $Q$ .

**Universality** Despite the great variety of physical systems that exhibit equilibrium second-order phase transitions, their critical behaviors, characterized by a set of critical exponents, fall into a small number of universality classes that only depend on the symmetry of the order parameter and the space dimension. Critical behavior is universal in the sense that it does not depend upon details whose characteristic size is much less than the correlation length, such as lattice structure, range of interactions (as long as this range is finite), spin length, and so on. Since time is involved in nonequilibrium critical phenomena, universality classes are expected to offer a richer variety.

**Cellular automaton** A cellular automaton is a fully discrete dynamical system. It consists of a regular finite-dimensional lattice of cells, each one in a state belonging to a finite set of states. The state of each cell evolves in discrete time steps. At time  $t$  the state of a given cell is a function of the states of a finite number of neighboring cells at time  $t - 1$ . The neighborhood of a given cell may or may not include the cell itself. All

cells evolve according to the same evolution rule which may be either deterministic or probabilistic.

**Model** A model is a simplified mathematical representation of a system. In the actual system, although many features are likely to be important, not all of them, however, should be included in the model. Only a few relevant features which are thought to play an essential role in the interpretation of the observed phenomena should be retained. If it captures the key elements of a complex system, a simple model, may elicit highly relevant questions.

**Mean-field approximation** In a many-agent system the mean-field approximation is a first attempt to understand the behavior of the system. Although rather crude, it is often useful. Ignoring space correlations between the agents and replacing local interactions by uniform long-range ones, the mean-field approximation only deals with average quantities. Historically, under the name “molecular field theory”, it was first used by Pierre Weiss (1869–1940) in 1907 to build up the first simple theory of the para-ferromagnetic second-order phase transition.

## Definition of the Subject

Phase transitions in cellular automata are non-equilibrium phase transitions observed in probabilistic cellular automata with absorbing states [1,2], that is, states that can be reached by the dynamics but cannot be left. Directed percolation is one of the most studied example of a non-equilibrium phase transition but many other examples such as epidemic or traffic models have attracted, since the 1990s, a considerable degree of attention.

## Introduction

In this article, the notion of critical behavior will often play an important role. Critical behavior manifests itself in many-component systems and is characteristic of a cooperative behavior of the various components. This notion has been introduced in statistical physics for many-body systems such as ferromagnetic materials, alloys, or liquid helium, that exhibit second-order phase transitions; that is, a phase change as a function of a tuning parameter, such as the temperature. A ferromagnetic material (i. e., a system having a spontaneous nonzero magnetization), becomes, as its temperature is (in general) increased, paramagnetic (i. e., its magnetization in the absence of an external magnetic field is equal to zero). In an ordered alloy such as  $\beta$ -brass – a 50 % copper and 50 % zinc alloy – the atoms of copper and zinc are located on two identical sublattices, one sublattice containing more copper and the other more zinc. As the temperature is increased, the alloy

becomes disordered (i.e., both sublattices contain equal fractions of copper and zinc). Liquid helium, which behaves as an ordinary liquid at temperatures above 2.19 K, becomes superfluid at lower temperatures. The temperature at which these phase transitions occur is called the critical temperature, and the system at the critical temperature – more generally at the critical point – is said to be in a critical state.

At the critical point, physical quantities, such as entropy, volume, or magnetization, that are first derivatives of the Gibbs free energy are continuous, in contrast with second-order derivatives, such as the specific heat or the magnetic susceptibility, which are singular. These singular behaviors reflect the long-range nature of the correlations in the vicinity of the critical point.

Close to a second-order phase transition, correlation functions of fluctuating quantities (such as spins in the case of a para-ferromagnetic phase transition) at two different points decrease exponentially with a characteristic correlation length  $\xi$ . As the temperature  $T$  approaches the critical temperature  $T_c$ ,  $\xi$  diverges as  $(T - T_c)^{-\nu}$  if  $T > T_c$  and  $(T - T_c)^{-\nu'}$  if  $T < T_c$ .

This cooperative effect is characteristic of criticality. It implies that certain physical quantities either vanish or diverge as powers of  $|T - T_c|$  as  $T$  approaches  $T_c$ . Today, when a many-agent system displays a power-law behavior for some observable, most researchers agree that this is a sign of some cooperative effect and a manifestation of the system complexity [3].

Despite the great variety of physical systems that exhibit second-order phase transitions, their critical behaviors, characterized by a set of critical exponents, fall into a small number of universality classes that only depend on the symmetry of the order parameter (such as the magnetization for a ferromagnet) and space dimension. Critical behavior is universal in the sense that it does not depend upon details whose characteristic size is much less than the correlation length, such as lattice structure, range of interactions (as long as this range is finite), spin length, and so on. Moreover, for a given second-order phase transition, one needs to know only a rather small number of critical exponents to determine all other exponents. For instance, in the case of a para-ferromagnetic second-order phase transition, the specific heat at constant magnetic field  $C_B$  diverges as  $(T - T_c)^{-\alpha}$  if  $T > T_c$  and  $(T - T_c)^{-\alpha'}$  if  $T < T_c$ , the magnetization  $M$ , which is identically equal to zero for  $T > T_c$ , goes to zero as  $(T_c - T)^\beta$  if  $T < T_c$ , and the isothermal susceptibility  $\chi_T$  diverges as  $(T - T_c)^{-\gamma}$  if  $T > T_c$  and  $(T - T_c)^{-\gamma'}$  if  $T < T_c$ . If we assume that the free energy  $F$ , close to the critical point, is a generalized homogeneous function of

$T - T_c$  and  $M$ , that is, a function satisfying, for all values of  $\lambda$ , the relation

$$F(\lambda(T - T_c), \lambda^\beta M) \equiv \lambda^{2-\alpha} F(T - T_c, M).$$

Choosing  $\lambda = 1/|T - T_c|$ , we can write  $F(T - T_c, M)$  under the form

$$F(T - T_c, M) = |T - T_c|^{2-\alpha} f\left(\frac{M}{|T - T_c|^\beta}\right),$$

where  $f$  is a function of only one variable. From this expression it can be shown (see [4]) that the critical exponents satisfy the following so-called scaling relations

$$\alpha = \alpha', \quad \gamma = \gamma', \quad \text{and} \quad \alpha + 2\beta + \gamma = 2.$$

An important distinguishing feature of the power-law behavior of physical quantities in the neighborhood of a critical point is that these quantities have no intrinsic scale. The function  $x \mapsto \exp(-x/\xi)$  has an intrinsic scale  $\xi$ , whereas the function  $x \mapsto x^a$  has no intrinsic scale: power laws are self-similar.

Quantities exhibiting a power-law behavior have been observed in a variety of disciplines ranging from linguistics and geography to medicine and economics. As mentioned above, the emergence of such a behavior is regarded as the signature of a collective mechanism.

Second-order phase transitions are always associated with a broken symmetry. That is, the symmetry group of the ordered phase (the phase characterized by a nonzero value of the order parameter) is a subgroup of the disordered phase (the phase characterized by an order parameter identically equal to zero). To an order parameter, we can always associate a symmetry-breaking field. In the presence of such a field, the order parameter has a nonzero value, and, in this case, the system cannot exhibit a second-order phase transition. In the case of an ideally simple para-ferromagnetic phase transition, the paramagnetic phase is invariant under the tridimensional rotation group whereas the ferromagnetic phase is no more invariant under that group. The corresponding broken symmetry is characterized by the nonzero value of the vector magnetization  $\mathbf{M}$  which plays the role of the order parameter, and the symmetry-breaking field is the magnetic field  $\mathbf{B}$  intensive conjugate parameter of  $\mathbf{M}$ .

The nature of the broken symmetry is not always obvious as, for instance, in the case of the normal-superfluid or normal-superconductor phase transitions, but it does exist (see [5]).

Depending upon the nature of the order parameter, a second-order phase transition exists only above a critical space dimensionality, called the lower critical space

dimension. Critical exponents, which depend upon space dimensionality, take their mean-field values (i. e., when space dependence and correlations are ignored) above another critical space dimensionality, called the upper critical space dimension. In the case of the Ising model, the lower and upper critical space dimensions are, respectively, equal to 1 and 4.

In what follows we focus on phase transitions in cellular automata. Deterministic one-dimensional cellular automaton rules are defined as follows. Let  $Q$  denote the finite set of integers  $\{0, 1, \dots, m\}$  and  $s(i, t) \in Q$  represent the state at site  $i \in \mathbb{Z}$  and time  $t \in \mathbb{N}$ ; a local evolution rule is a map  $f: Q^{r_\ell + r_r + 1} \rightarrow Q$  such that

$$s(i, t+1) = f(s(i-r_\ell, t), s(i-\ell+1, t), \dots, s(i+r_r, t)), \quad (1)$$

where the integers  $r_\ell$  and  $r_r$  are, respectively, the left radius and right radius of the rule  $f$ ; if  $r_\ell = r_r = r$ ,  $r$  is called the radius of the rule. The local rule  $f$ , which is a function of  $n = r_\ell + r_r + 1$  arguments, is often said to be an  $n$ -input rule. The function  $S_t: i \mapsto s(i, t)$  is the state of the cellular automaton at time  $t$ ;  $S_t$  belongs to the set  $Q^{\mathbb{Z}}$  of all configurations. Since the state  $S_{t+1}$  at  $t+1$  is entirely determined by the state  $S_t$  at time  $t$  and the local rule  $f$ , there exists a unique mapping  $F_f: S \rightarrow S$ , such that

$$S_{t+1} = F_f(S_t), \quad (2)$$

called the cellular automaton global rule or the cellular automaton evolution operator induced by the local rule  $f$ .

Most models presented in this article will be formulated in terms of probabilistic cellular automata. In this case, the image by the evolution rule of any  $(r_\ell + r_r + 1)$ -block, is a discrete random variable with values in  $Q$ .

The simplest cellular automata are the so-called elementary cellular automata in which the finite set of states is  $Q = \{0, 1\}$  and the rule's radii are  $r_\ell = r_r = 1$ . Sites in a nonzero state are sometimes said to be active. It is easy to verify that there exist  $2^{2^3} = 256$  different elementary cellular automaton local rules  $f: \{0, 1\}^3 \rightarrow \{0, 1\}$ . The local rule of an elementary cellular automaton can be specified by its look-up table, giving the image of each of the eight three-site neighborhoods. That is, any sequence of eight binary digits specifies an elementary cellular automaton rule. Here is an example:

111	110	101	100	011	010	001	000
1	0	1	1	1	0	0	0

Following Wolfram [6], a code number may be associated with each cellular automaton rule. If  $Q = \{0, 1\}$ , this

code number is the decimal value of the binary sequence of images. For instance, the code number of the rule above is 184 since

$$10111000_2 = 2^7 + 2^5 + 2^4 + 2^3 = 184_{10}.$$

More generally, the code number  $N(f)$  of a one-dimensional  $|Q|$ -state  $n$ -input cellular automaton rule  $f$  is defined by

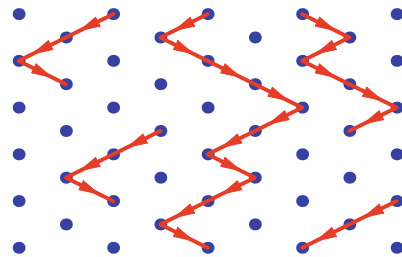
$$N(f) = \sum_{(x_1, x_2, \dots, x_n) \in Q^n} f(x_1, x_2, \dots, x_n) \cdot |Q|^{n-1}x_1 + |Q|^{n-2}x_2 + \dots + |Q|^0x_n.$$

### The Domany-Kinzel Cellular Automaton

Directed percolation refers to lattice models that mimic coffee brewing, that is, causing water to pass through a bed of ground coffee. Consider the square lattice represented in Fig. 1 in which open bonds are randomly distributed with a probability  $p$ . In contrast with the usual bond percolation problem, here bonds are directed downwards, as indicated by the arrows.

If we imagine a fluid flowing downwards from wet sites in the first row, one problem is to find the probability  $P(p)$  that, following directed open bonds, the fluid will reach sites on an infinitely distant last row. There clearly exists a threshold value  $p_c$  above which  $P(p)$  is nonzero (see [7]). If the downwards direction is considered to be the time direction, the directed bond percolation process may be viewed as the evolution of a two-input one-dimensional cellular automaton rule  $f$  such that

$$s(i, t+1) = f(s(i, t), s(i+1, t)) = \begin{cases} 0, & \text{if } s(i, t) + s(i+1, t) = 0, \\ 1, & \text{with probability } p, \text{ if } s(i, t) + s(i+1, t) = 1, \\ 1, & \text{with probability } 1 - (1-p)^2, \\ & \text{if } s(i, t) + s(i+1, t) = 2. \end{cases}$$



Phase Transitions in Cellular Automata, Figure 1

A configuration of directed bond percolation on a square lattice

The density of active (wet) sites  $\rho$ , which is equal to  $P(p)$ , plays the role the order parameter of the second-order phase transition. Using the image of the flowing fluid, it can be shown that there exists a directed bond percolation threshold (or a directed bond percolation probability)  $p_c^{\text{DBP}}$  above which the fluid has a nonzero probability of reaching an infinitely distant last row. In the vicinity of the critical probability  $p_c^{\text{DBP}}$ , if  $\xi_{\parallel}$  and  $\xi_{\perp}$  denote, respectively, the correlation length in the flow direction and perpendicular to it, we have

$$\xi_{\parallel} \sim \xi_{\perp}^{\theta} \sim (p - p_c^{\text{DBP}})^{-\nu_{\parallel}} \sim (p - p_c^{\text{DBP}})^{-\nu_{\perp}\theta},$$

where  $\theta$  is the anisotropy exponent and  $\nu_{\parallel}$  and  $\nu_{\perp}$  the correlation length exponents in the longitudinal and transverse directions respectively. Using the finite-size renormalization group technique Kinzel and Yeomans, assuming free boundary conditions, found [8]

$$p_c^{\text{DBP}} = 0.644 \pm 0.001 \quad \theta = 1.582 \pm 0.001 \\ \nu_{\parallel} = 1.739 \pm 0.002 \quad \nu_{\perp} = 1.099 \pm 0.001.$$

A cellular automaton  $n$ -input rule is said to be totalistic if it only depends upon the sum of the  $n$  inputs. The most general two-input one-dimensional totalistic probabilistic cellular automaton rule, called the Domany–Kinzel cellular automaton rule [9], may be written

$$s(i, t+1) = f(s(i, t), s(i+1, t)) \\ = \begin{cases} 0, & \text{if } s(i, t) + s(i+1, t) = 0, \\ 1, & \text{with probability } p_1, \text{ if } s(i, t) + s(i+1, t) = 1, \\ 1, & \text{with probability } p_2, \text{ if } s(i, t) + s(i+1, t) = 2. \end{cases}$$

Directed bond percolation corresponds to  $p_1 = p$  and  $p_2 = 2p - p^2$ . The case  $p_1 = p_2 = p$  is also interesting; it describes the *directed site percolation* process. In this case, numerical simulations show that the directed site percolation probability  $p_c^{\text{DSP}} = 0.7058 \pm 0.0005$  (values of the critical exponents characterizing the singular behavior close to the critical probability  $p_c^{\text{SDP}}$  can be found in [8]).

Domany and Kinzel showed that, in the infinite time limit, there exist two phases, an active phase in which a macroscopic fraction of all sites are occupied (state value equal to 1) and a phase in which all sites become empty (state value equal to 0). The domains of existence of these two phases in the  $(p_1, p_2)$ -plane are separated by a second-order phase transition line (see [9]). Along this line, all phase transitions belong to the same universality class

characterized by a critical exponent  $\beta = 0.273 \pm 0.002$ , characterizing the singular behavior of the order parameter in the vicinity of the critical temperature.

A few years later, Martins, Verona de Resende, Tsallis, and de Magalhães [10] considered a generalized version of the Domany–Kinzel cellular automaton whose evolution rule is given by

$$s(i, t+1) = f(s(i, t), s(i+1, t)) \\ = \begin{cases} 0, & \text{if } s(i, t) = 0 \text{ and } s(i+1, t) = 0, \\ 1, & \text{with probability } p_1, \\ & \text{if } s(i, t) = 0 \text{ and } s(i+1, t) = 1, \\ 1, & \text{with probability } p_3, \\ & \text{if } s(i, t) = 1 \text{ and } s(i+1, t) = 0, \\ 1, & \text{with probability } p_2, \\ & \text{if } s(i, t) = 1 \text{ and } s(i+1, t) = 1, \end{cases}$$

that is, a two-input one-dimensional probabilistic cellular automaton rule which is no more totalistic. In the three-dimensional  $(p_1, p_2, p_3)$ -phase space they found, in the infinite time limit, a new chaotic phase. The corresponding phase transition does not belong to the same universality class as the Domany–Kinzel phase transition. In particular, the critical exponent  $\beta = 0.5 \pm 0.02$ . The boundaries between the three phases of the Domany–Kinzel probabilistic cellular automaton have been determined with high accuracy in [11]. For a renormalization group approach of the Domany–Kinzel cellular automaton refer to [12].

A richer phase diagram of a simple three-input one-dimensional totalistic cellular automaton has been studied by Bagnoli, Boccara, and Rechtman [13]. These authors studied the phase diagram and the critical behavior of the one-dimensional radius-1 totalistic probabilistic cellular automaton whose evolution rule is defined as follows. If  $s(i, t)$  denotes the state of the  $i$ th cell at time  $t$ , then

$$s(i, t+1) \\ = \begin{cases} 0, & \text{if } s(i-1, t) + s(i, t) + s(i+1, t) = 0, \\ X_1, & \text{if } s(i-1, t) + s(i, t) + s(i+1, t) = 1, \\ X_2, & \text{if } s(i-1, t) + s(i, t) + s(i+1, t) = 2, \\ 1, & \text{if } s(i-1, t) + s(i, t) + s(i+1, t) = 3, \end{cases}$$

where  $X_j$  ( $j = 1, 2$ ) is a Bernoulli random variable equal to 1 with probability  $p_j$ , and to 0 with probability  $1 - p_j$ . In the  $(p_1, p_2)$ -plane, the line  $p_1 + p_2 = 1$  is a symmetry axis of the phase diagram. The evolution rule implies that states in which the cells are either all empty or all occupied

are absorbing. There exists a first-order phase transition between these two phases along the line  $p_1 + p_2 = 1$  ending at a multicritical point where two second-order phase transition lines meet. These second-order transition lines separate the absorbing states mentioned above from a stable phase having a density  $\rho$  of occupied sites such that  $0 < \rho < 1$ , (i.e.,  $\rho \neq 0$  and  $\rho \neq 1$ ). The two second-order phase transitions belong to the universality class of the directed percolation phase transition. Finally there exists a chaotic phase, located in the neighborhood of the point  $(p_1, p_2) = (1, 0)$  in the  $(p_1, p_2)$ -phase plane of the type discovered by Martins, Verona de Resende, Tsallis, and de Magalhães.

### Car Traffic Models

Vehicular traffic can be treated as a system of interacting particles driven far from equilibrium. The particle-hopping model describes car traffic in terms of probabilistic cellular automata. An interesting model of this type was proposed by Nagel and Schreckenberg [14]. These authors consider a finite lattice of length  $L$  with periodic boundary conditions. A finite one-dimensional cellular automaton of length  $L$  is said to satisfy periodic boundary conditions if the set of vertices is  $\mathbb{Z}_L$ , that is, the set of integers modulo  $L$ . In the case of a one-lane traffic model, these conditions are equivalent to assuming that cars are moving on a circular one-lane highway with neither entries nor exits. Each cell is either empty (i.e., in state  $e$ ) or occupied by a car (i.e., in state  $v$ ), where  $v = 0, 1, \dots, v_{\max}$  denotes the

car velocity (cars are moving to the right). If  $d_i$  is the distance between cars  $i$  and  $i + 1$ , car velocities are updated in parallel according to the following subrules.

$$v_i(t + \frac{1}{2}) = \min(v_i(t) + 1, d_i(t) - 1, v_{\max})$$

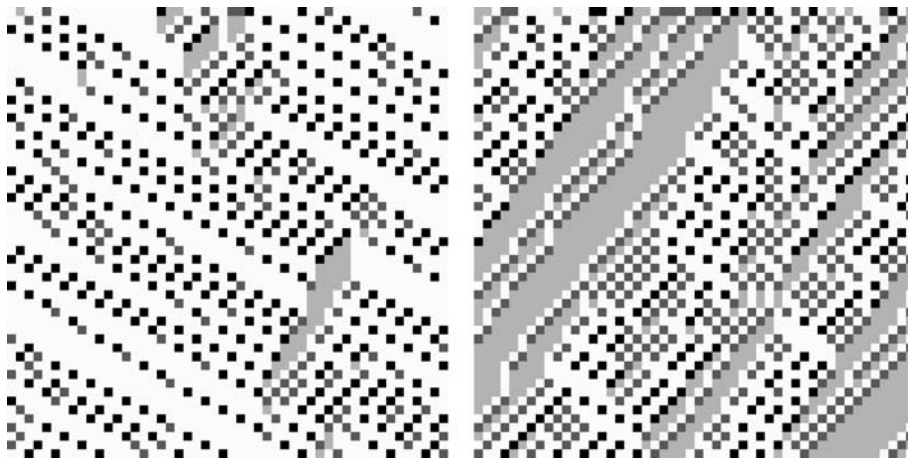
$$v_i(t + 1) = \begin{cases} \max(v_i(t + \frac{1}{2}) - 1, 0) , & \text{with probability } p, \\ v_i(t + \frac{1}{2}) , & \text{with probability } 1 - p , \end{cases} \quad (3)$$

where  $v_i(t)$  is the velocity of car  $i$  at time  $t$ . Then, if  $x_i(t)$  is the position of car  $i$  at time  $t$ , cars are moving according to the rule

$$x_i(t + 1) = x_i(t) + v_i(t + 1) .$$

That is, at each time step, each car increases its speed by one unit (acceleration  $a = 1$ ), respecting the safety distance and the speed limit. The model also includes noise: with a probability  $p$ , each car decreases its speed by one unit. Although rather simple, the model exhibits features observed in real highway traffic, that is, with increasing vehicle density, it shows a phase transition from laminar traffic flow to start-stop waves as illustrated in Fig. 2.

In order to understand, in the case of a second-order phase transition in a highway car traffic cellular automaton model, the nature of the order parameter, show how it is related to symmetry-breaking, determine the symmetry-breaking field conjugate to the order parameter, define the



Phase Transitions in Cellular Automata, Figure 2

First 50 iterations of the Nagel–Schreckenberg probabilistic cellular automaton traffic flow model. The initial configuration is random with a density equal to 0.2 in the *left* figure and 0.5 in the *right* one. In both cases  $v_{\max} = 2$  and  $p = 0.2$ . The number of lattice sites is equal to 50. Empty cells are white while cells occupied by a car with velocity  $v$  equal to 0, 1, and 2 have *darker shades of gray*. Time increases downwards



analogue of the susceptibility, study the critical behavior, and find scaling laws, Boccaro and Fuks studied in details a deterministic version of the Nagel–Schreckenberg highway traffic model [15].

If  $p = 0$ , the Nagel–Schreckenberg model is deterministic and the average velocity over the whole lattice is exactly given by

$$\langle v \rangle = \min \left( v_{\max}, \frac{1}{\rho} - 1 \right). \quad (4)$$

This expression shows that, below a critical car density

$$\rho_c = 1/(v_{\max} + 1),$$

all cars move with a velocity equal to  $v_{\max}$ , while above  $\rho_c$ , the average velocity is less than  $v_{\max}$ .

To further simplify the Nagel–Schreckenberg model, we assume that the acceleration, which is equal to 1 in the Nagel–Schreckenberg model, has the largest possible value (less or equal to  $v_{\max}$ ) as in the Fukui–Ishibashi model [16]. That is, in our model, we just replace (3) by

$$v_i(t + 1/2) = \min(d_i(t) - 1, v_{\max}) \quad (5)$$

Deterministic cellular automaton rules modeling traffic flow on one-lane highways are number-conserving (i. e.,  $\rho = \text{constant}$ ). Limit sets of number-conserving cellular automata have, in most cases, a very simple structure and, these limit sets are reached after a number of time steps proportional to the lattice size [17,18,19] as illustrated in Fig. 3.

If  $\rho \leq \rho_c$ , any configuration in the limit set consists of “perfect tiles” of  $v_{\max} + 1$  cells as shown below

$v_{\max}$	$e$	$\dots$	$e$	$e$
------------	-----	---------	-----	-----

in a sea of cells in state  $e$ .

If  $\rho > \rho_c$ , a configuration belonging to the limit set only consists of a mixture of tiles containing  $v + 1$  cells of the type

$v$	$e$	$\dots$	$e$	$e$
-----	-----	---------	-----	-----

where  $v = 0, 1, \dots, v_{\max}$ . For  $v < v_{\max}$ , all these are said to be “defective.” If  $\{\rho_v \mid v = 0, 1, 2, \dots, v_{\max}\}$  is the velocities distribution, we have

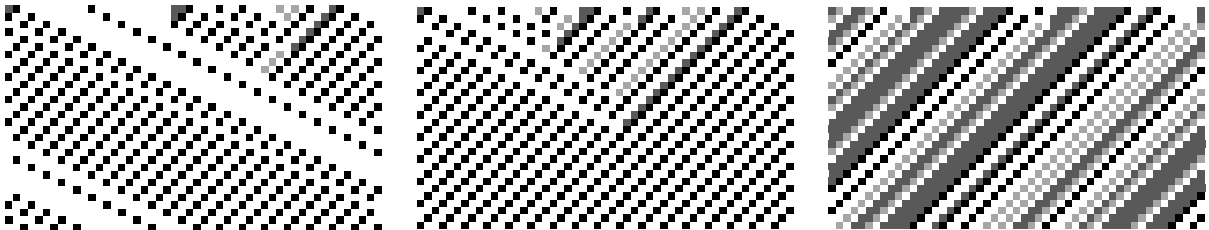
$$\begin{aligned} \rho &= \sum_{v=0}^{v_{\max}} \rho_v \\ 1 &= \sum_{v=0}^{v_{\max}} (v + 1) \rho_v \\ \langle v \rangle &= \frac{1}{\rho} \left( \sum_{v=0}^{v_{\max}} v \rho_v \right). \end{aligned}$$

Note that Relation (4) is a simple consequence of these relations.

If we introduce random braking, then, even at low density, some tiles become defective, which causes the average velocity to be less than  $v_{\max}$ . The random-braking parameter  $p$ , which is an essential ingredient of all cellular automaton traffic flow model, can, therefore, be viewed as a symmetry-breaking field, and the order parameter, conjugate to that field is

$$m = v_{\max} - \langle v \rangle \quad (6)$$

This point of view implies that the phase transition characterized by  $m$  will be smeared out in the presence of random braking as a para-ferromagnetic phase transition in the presence of a magnetic field.



Phase Transitions in Cellular Automata, Figure 3

First 30 iterations of the deterministic cellular automaton traffic model for  $v_{\max} = 2$ . The critical density  $\rho_c$  is equal to  $1/3$ . Initial configurations are random with a density exactly equal to 0.26 in the left figure,  $1/3$  in the central one, and 0.6 in the right one. The number of lattice sites is equal to 50 in the left and right figures and to 51 in the central figure. Empty cells are white whereas cells occupied by a car with velocity  $v$  equal to either 0, 1, or 2 have darker shades of gray. Time increases downwards. Note that for  $\rho \leq 1/3$ , all cars move at the speed limit  $v_{\max} = 2$ .

From (6) and (4), it follows that, for  $p = 0$ ,

$$m = \begin{cases} 0 & \text{if } \rho \leq \rho_c, \\ \frac{\rho - \rho_c}{\rho \rho_c} & \text{otherwise.} \end{cases} \quad (7)$$

The critical exponent  $\beta$  is, therefore, equal to 1.

The other critical exponents cannot be found exactly but can be determined using numerical simulations. In our simulations, we took  $v_{\max} = 2$ , used a lattice size equal to 1000 and we averaged our results over 1000 runs of 1000 iterations. For  $p = 0.0005$  we took a lattice of 10 000 sites and averaged 500 runs of 10 000 iterations [15].

The susceptibility  $\chi_\rho$  at constant  $\rho$ , defined by

$$\chi_\rho = \lim_{p \rightarrow 0} \frac{\partial m}{\partial p}. \quad (8)$$

In the limit  $p \rightarrow 0$ , the susceptibility diverges as  $(\rho_c - \rho)^{-\gamma}$  for  $\rho < \rho_c$ , and as  $(\rho - \rho_c)^{-\gamma'}$  for  $\rho > \rho_c$ . Our simulations yield

$$\gamma = 0.86 \pm 0.05 \quad \text{and} \quad \gamma' = 0.94 \pm 0.05.$$

Another exponent of interest is  $\delta$ . It characterizes the behavior of  $m$  as a power of  $p$  for  $\rho = \rho_c$ . Here again we have determined the value of

$$\lim_{p \rightarrow 0} \frac{m(\rho_c, 0) - m(\rho_c, p)}{p}$$

using numerical simulations. Our result is

$$1/\delta = 0.53 \pm 0.02$$

It is interesting to note that the values

$$\beta = 1, \quad \gamma \approx 1, \quad \delta \approx 2$$

obtained for the critical exponents are found in equilibrium statistical physics in the case of second-order phase transitions characterized by nonnegative order parameters above the upper critical dimensionality.

Close to the phase transition point, critical exponents obey scaling relations. If we assume that, in the vicinity of the critical point ( $\rho = \rho_c$ ,  $p = 0$ ), the order parameter  $m$  is a generalized homogeneous function of  $\rho - \rho_c$  and  $p$  of the form

$$m = |\rho - \rho_c|^\beta f\left(\frac{p}{|\rho - \rho_c|^{\beta\delta}}\right), \quad (9)$$

where the function  $f$  is such that  $f(0) \neq 0$ , then, differentiating  $f$  with respect to  $p$  and taking the limit  $p \rightarrow 0$ , we readily obtain

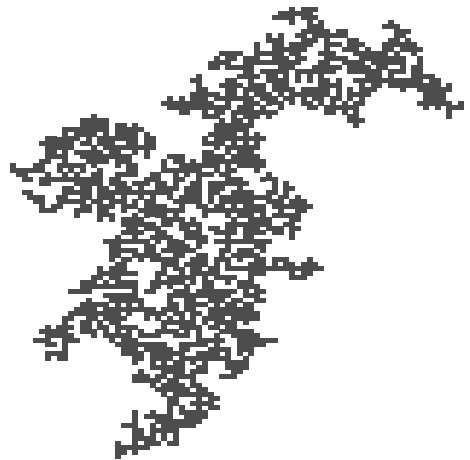
$$\gamma = \gamma' = (\delta - 1)\beta, \quad (10)$$

in agreement with our numerical simulations.

Boccara [20] has shown that this highway traffic flow model satisfies, with other deterministic traffic flow models, a variational principle.

## Epidemic Models

The general epidemic process (see Fig. 4) describes, according to Grassberger [21,22], who studied its critical properties, “the essential features of a vast number of population growth phenomena.” In its simplest version the growth process can be described as follows. Initially the cluster consists of the seed site located at the origin. At the next time step, a nearest-neighboring site is randomly chosen. This site is either added to the cluster with a probability  $p$  or rejected with a probability  $1 - p$ . At all subsequent time steps, the same process is repeated: a nearest-neighboring site of any site belonging to the cluster is selected at random, and it is either added to the cluster with a probability  $p$  or rejected with a probability  $1 - p$ . It is clear that there exists a critical probability  $p_c$  such that for  $p > p_c$ , the seed site has a nonzero probability of belonging to an infinite cluster. In order to determine the critical behavior of the general epidemic model, Grassberger performed extensive numerical simulations on a slightly different model that



Phase Transitions in Cellular Automata, Figure 4

Grassberger's general epidemic process. The cluster represents the spread of the epidemic to 3000 sites for  $p = 0.6$

belongs to the same universality class and whose static properties are identical to a bond percolation model. In this model, every lattice site of a two-dimensional square lattice is occupied by only one individual, who cannot move away from it. The individuals are either susceptible, infected, or immune. At each time step, every infected individual infects each nearest-neighbor susceptible site with a probability  $p$  and becomes immune with probability 1. For this model, the critical probability  $p_c$  is exactly equal to  $1/2$ . If, at time  $t = 0$ , all the sites of one edge of the lattice are infected, among other results, Grassberger found that, at  $p_c$ , the average number of immune sites per row parallel to the initial infected row increases as a function of time as  $t^x$ , where the critical exponent  $x$  is equal to  $0.807 \pm 0.01$  (for more detailed numerical results, refer to Grassberger [21]).

Epidemic models have a long history that started in 1927 with the publication by Kermack and McKendrick of their celebrated “threshold theorem” stating that the spread of a disease occurs only if the density of individuals susceptible to catch the disease is greater than a threshold value [23]. Here we shall only describe a model formulated in terms of cellular automata due to Boccara and Cheong [24,25] that exhibits a phase transition. In this so-called SIS epidemic model, individuals are divided into two disjoint groups:

1. *susceptible* individuals capable of contracting the disease and becoming infective, and
2. *infective* individuals capable of transmitting the disease to susceptibles.

If  $p_i$  denotes the probability for a susceptible to be infected and  $p_r$  the probability for an infective to recover and return to the susceptible group, the possible evolution of an individual may be represented by the following transfer diagram:

$$S \xrightarrow{p_i} I \xrightarrow{p_r} S.$$

In a two-dimensional cellular automaton model, with periodic boundary conditions, in which the sites are elements of the space  $\mathbb{Z}_L \times \mathbb{Z}_L$ , each site is either empty or occupied by a susceptible or an infective.

The spread of the disease is governed by the following rules.

1. Susceptible individuals become infected by contact (i. e., a susceptible may become infective with a probability  $p_i$  if, and only if, it is in the neighborhood of an infective). This hypothesis neglects incubation and la-

tent periods: an infected susceptible becomes immediately infective.

2. Infective individuals recover and become susceptible again with a probability  $p_r$ . This assumption states that recovery is equally likely among infective individuals but does not take into account the length of time the individual has been infective.
3. The time unit is the time step. During one time step, the two preceding rules are applied synchronously, and the individuals move on the lattice according to a specific rule.
4. An individual selected at random performs a move. That is, a site occupied by an individual is selected at random and swapped with another site (either empty or occupied) also selected at random. If the second site is a nearest neighbor of the first site, the resulting move of the individual is said to be short-range whereas it is said to be long-range if it is any site of the lattice. This operation is repeated  $\lfloor m \times \rho \times L^2 \rfloor$  times, where  $m$  is a positive real number called the degree of mixing and  $\rho$  the density of occupied sites at time  $t$ . When two occupied sites are swapped the move is not effective;  $m$  is therefore the average number of tentative moves. The notation  $\lfloor x \rfloor$  represents the largest integer less than or equal to  $x$ .

The model assumes that the population is closed; it therefore ignores births, deaths by other causes, immigrations, or emigrations.

The critical behavior of this model has been studied by means of numerical simulations [24]. The total density of individuals was equal to 0.6, slightly above the site percolation threshold in two dimensions for the square lattice in order to be able to observe cooperative effects. Most simulations were performed on a  $100 \times 100$  lattice and some on a  $200 \times 200$  lattice to check possible size effects.

In the case of short-range moves, for given values of  $p_r$  and  $m$ , there exists a critical value  $p_i^c$  of the probability for a susceptible to become infected. At this transition point, the stationary density of infective individuals  $I_\infty(m)$  behaves as  $(p_i - p_i^c)^\beta$ . When  $m = 0$  (i. e., if individuals do not move), the exponent  $\beta$  is close to 0.6, which is the value for the two-dimensional directed percolation.

For a given value of  $p_r$ , the variations of  $\beta$  and  $p_i^c$  as functions of  $m$  are found to exhibit two regimes reminiscent of crossover phenomena. In the small  $m$  regime (i. e., for  $m \lesssim 10$ ),  $p_i^c$  and particularly  $\beta$  have their  $m = 0$  values. In the large  $m$  regime (i. e., for  $m \gtrsim 300$ ),  $p_i^c$  and  $\beta$  have their mean-field values. In agreement with what is known in phase transition theory, the exponent  $\beta$  does not

seem to depend upon  $p_r$ ; i.e., its value does not change along the second-order transition line.

For given values of  $p_i$  and  $p_r$ , the asymptotic behaviors of the stationary density of infective individuals  $I_\infty(m)$  for small and large values of  $m$  may be characterized by the exponents

$$\alpha_0 = \lim_{m \rightarrow 0} \frac{\log(I_\infty(m) - I_\infty(0))}{\log m},$$

$$\alpha_\infty = \lim_{m \rightarrow \infty} \frac{\log(I_\infty(\infty) - I_\infty(m))}{\log m}.$$

It is found that  $\alpha_0 = 0.177 \pm 0.15$  and  $\alpha_\infty = -0.945 \pm 0.065$ .

The fact that  $\alpha_0$  is rather small shows the importance of motion in the spread of a disease. The stationary number of infective individuals increases dramatically when the individuals start to move. In other words, the response  $\partial I_\infty(m)/\partial m$  of the stationary density  $I_\infty(m)$  to the degree of mixing  $m$  tends to  $\infty$  when  $m$  tends to 0.

In the case of long-range moves, for a fixed value of  $p_r$ , the variations of  $p_i^c$  and  $\beta$  are very different from those for short-range moves. Whereas for short-range moves  $\beta$  and  $p_i^c$  do not vary in the small- $m$  regime, for long-range moves, on the contrary, the derivatives of  $\beta$  and  $p_i^c$  with respect to  $m$  tend to  $\infty$  as  $m$  tends to 0. For small  $m$ , the asymptotic behaviors of  $\beta$  and  $p_i^c$  may therefore be characterized by the exponents

$$\alpha_\beta = \lim_{m \rightarrow 0} \frac{\log(\beta(m) - \beta(0))}{\log m},$$

$$\alpha_{p_i^c} = \lim_{m \rightarrow 0} \frac{\log(p_i^c(m) - p_i^c(0))}{\log m}.$$

Both exponents are found to be close to 0.5.

## Future Directions

In the article we tried to focus on the essential characteristics of phase transitions in cellular automata illustrating our discussion with representative examples. In this section we briefly present other examples and list some articles that go deeper into the examples we chose to discuss.

One of the earliest example of a phase transition in cellular automata has been studied in 1984 by Grassberger et al. [26] who showed that the spatial patterns of two probabilistic cellular automata exhibit a transition from stability to instability of kinks between ordered states. As a function of the probability  $p$  the cellular automaton rules 94 and 50

for  $p = 0$  are continuously modified to become, for  $p = 1$ , rules 22 and 122 respectively. In both cases the authors determined the critical probabilities and a few critical exponents.

In 1985 Kinzel [27] investigated phase transitions of probabilistic two-state three-input one-dimensional cellular automata with absorbing states. Using a transfer matrix technique, he determined phase diagrams and critical exponents. He also studied a special three-state probabilistic cellular automaton that could be mapped onto a two-state cellular automaton modeling the spread of an epidemic taking into account immunization. For a field theoretic treatment of this epidemic model, see Cardy [28].

A particular class of probabilistic two-dimensional two-state cellular automata defined on  $\mathbb{Z}_L \times \mathbb{Z}_L$  (i.e., on a square lattice of size  $L$  with periodic boundary conditions) with nearest-neighbor interactions were investigated by Kaneko and Akutsu [29] who considered the evolution rule

$$s(t+1, i, j) = \begin{cases} f(\sigma(t, i, j), s(t, i, j)), & \text{with probability } 1-p, \\ 1-f(\sigma(t, i, j), s(t, i, j)), & \text{with probability } p, \end{cases}$$

where  $s(t, i, j)$  denotes the state at time  $t$  of the cell at site  $(i, j)$ ,  $\sigma(t, i, j) = s(t, i, j-1) + s(t, i, j+1) + s(t, i-1, j) + s(t, i+1, j)$ , and  $f$  is a function of two variables which takes the value 0 or 1. For small values of the probability  $p$ , they found a rich variety of phases.

In epidemic models we stressed the importance of individuals' motion. In our cellular automaton models, motion was modeled by a site-exchange process. That is, we considered cellular automata whose evolution rule consists of two subrules; the first one, applied synchronously, is a usual cellular automaton rule, whereas the second, applied sequentially, is a local or nonlocal exchange of two site values [30,31]. The evolution of a probabilistic site-exchange cellular automaton depends, therefore, upon two parameters, the probability  $p$  characterizing the probabilistic cellular automaton rule, and the degree of mixing  $m$  resulting from the exchange process (for the precise definition of  $m$ , refer above). Depending upon the values of these two parameters, the system exhibits a second order phase transition characterized by a nonnegative order parameter, whose role is played by the stationary density of occupied sites. When  $m$  is very large, the correlations created by the application of the probabilistic cellular automaton rule are destroyed and, as expected, the behavior of the system is then correctly described by

a mean-field-type approximation. According to whether the exchange of site values is local or nonlocal, the critical behavior is qualitatively different as  $m$  varies [32]. In [33], the authors found in the  $(p, m)$ -plane that, along the transition line, increasing  $m$ , the order of the phase transition changes from second- to first-order at a tricritical point characterized by a different critical behavior.

The Domany–Kinzel cellular automaton is a popular toy model that has attracted a lot of attention. A significant number of papers devoted to its study have been published. It has been shown that the qualitative features of the phase diagram, including the new phase found by Martins, Verona de Resende, Tsallis, and de Magalhães can be predicted analytically by going one-step beyond the mean-field approximation [34]. Although there is a clear numerical evidence that the critical behavior along the critical line found by Domany and Kinzel is that of directed percolation [1,35], this is not the case at terminal point  $(1/2, 1)$  [36,37]. The Domany–Kinzel model has also been used to illustrate the breakdown of universality in transitions to spatiotemporal chaos [38] and, recently, a few limit theorems have been rigorously established [39].

Since the early 1990s, traffic problems have drawn considerable attention and a great number of cellular automaton models of traffic flow have been proposed to deal with many diverse situations such as the existence of a jamming transition in two dimensions [40], the influence of two-level crossings on traffic jams [41], the effect of traffic accidents on the jamming transition [42], the crossing of two roads [43], the existence of a roadblock and the resulting number of stopped cars [44,45], and building up models of city traffic [46]. In the case of deterministic cellular automaton traffic flow models generalizing the cellular automaton rule 184 [47], that is, models in which the maximum speed is greater than 1, Fukš [48] has been able to derive exactly the flow diagram, that is, the graph of the car flow as a function of the car density. There exist also quite a few cellular automaton models of pedestrian traffic that exhibit phase transitions similar to those found in car traffic [49,50]. A very simple cellular automaton pedestrian model exhibiting self-organized motion in a multilane passageway with pedestrians moving in opposite directions is described in [3] page 204. For a detailed recent review on the application of statistical physics to traffic see [51]. Concerning “realistic” traffic, there exists an agent-based simulation project at Los Alamos National Laboratory called TRANSIMS (TRANSPORTATION ANALYSIS and SIMULATION SYSTEM) [52] “capable of simulating the second-by-second movements of every

person and every vehicle through the transportation network of a large metropolitan area” [53].

## Bibliography

### Primary Literature

1. Ódor G (2004) Universality classes in nonequilibrium lattice systems. *Rev Mod Phys* 76:663–724
2. Hinrichsen H (2006) Non-equilibrium phase transitions. *Physica A* 369:1–28
3. Boccarra N (2004) *Modeling Complex Systems*. Springer, New York
4. Boccarra N (1976) *Symétries Brisées*. Hermann, Paris
5. Boccarra N (1972) On the microscopic formulation of Landau theory of phase transition. *Solid State Commun* 11:131–141
6. Wolfram S (1983) Statistical physics of cellular automata. *Rev Mod Phys* 55:601–644
7. Kinzel W (1983) Directed percolation in Percolation, Structures and Processes. *Ann Isr Phys Soc* 5:425–445
8. Kinzel W, Yeomans J (1981) Directed percolation: a finite-size scaling renormalization group approach. *J Phys A* 14: L163–L168
9. Domany E, Kinzel W (1984) Equivalence of cellular automata to Ising models and directed percolation. *Phys Rev Lett* 53: 311–314
10. Martins ML, Verona de Resende HF, Tsallis C, Magalhães ACN (1991) Evidence of a new phase in the Domany–Kinzel cellular automaton. *Phys Rev Lett* 66:2045–2047
11. Zebende GF, Penna TJP (1994) The Domany–Kinzel cellular automaton phase diagram. *J Stat Phys* 74:1274–1279
12. Tomé T, de Oliveira J (1997) Renormalization group of the Domany–Kinzel cellular automaton. *Phys Rev E* 55:4000–4004
13. Bagnoli F, Boccarra N, Rechtman R (2001) Nature of phase transitions in a probabilistic cellular automaton with two absorbing states. *Phys Rev E* 63:046 116
14. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I* 2:2221–2229
15. Boccarra N, Fukš H (2000) Critical behavior of a cellular automaton highway traffic model. *J Phys A: Math Gen* 33:3407–3415
16. Fukui M, Ishibashi Y (1996) Traffic flow in a 1D cellular automaton model including cars moving with high speed. *J Phys Soc Jpn* 65:1868–1870
17. Boccarra N, Nasser J, Roger M (1991) Particlelike structures and their interactions in spatiotemporal patterns generated by one-dimensional deterministic cellular-automaton rules. *Phys Rev A* 44:866–875
18. Boccarra N, Fukš H (1998) Cellular automaton rules conserving the number of active sites. *J Phys A: Math Gen* 31:6007–6018
19. Boccarra N, Fukš H (2002) Number-conserving cellular automaton rules. *Fundamenta Informaticae* 52:1–13
20. Boccarra N (2001) On the existence of a variational principle for deterministic cellular automaton models of highway traffic flow. *Int J Mod Phys C* 12:1–16
21. Grassberger P (1983) On the critical behavior of the general epidemic process and dynamical percolation. *Math Biosci* 63:157–172



22. Cardy JL, Grassberger P (1985) Epidemic models and percolation. *J Phys A: Math Gen* 18:L267–L271
23. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proc Royal Soc A* 115: 700–721
24. Boccara N, Cheong K (1993) Critical behaviour of a probabilistic automata network SIS model for the spread of an infectious disease in a population of moving individuals. *J Phys A: Math Gen* 26:3707–3717
25. Boccara N, Cheong K (1993) Automata network epidemic models. In: Boccara N, Goles E, Martínez S, Picco P (eds) *Cellular Automata and Cooperative Systems*. Kluwer, Dordrecht, pp 29–44
26. Grassberger P, Krause F, Von der Twer T (1984) A new type of kinetic critical phenomenon. *J Phys A: Math Gen* 17:L105–L109
27. Kinzel W (1985) Phase transitions in cellular automata. *Z Phys B* 58:229–244
28. Cardy J (1983) Field theoretic treatment of an epidemic process with immunisation. *J Phys A* 16:L709–L712
29. Kaneko K, Akutsu Y (1986) Phase transitions in two-dimensional stochastic cellular automata. *J Phys A: Math Gen* 19: L69–L75
30. Boccara N, Roger M (1993) Site-exchange cellular automata. In: Tirapegui E, Zeller W (eds) *Instabilities and Nonequilibrium Structures, IV*. Kluwer Dordrecht, pp 109–118
31. Boccara N, Roger M (1994) Some properties of local and non-local site-exchange deterministic cellular automata. *Int J Mod Phys C* 5:581–588
32. Boccara Nasser J, Roger M (1994) Critical behavior of a probabilistic local and nonlocal site-exchange cellular automaton. *Int J Mod Phys C* 5:537–545
33. Ódor G, Boccara N, Szabo G (1993) Phase-transition study of a one-dimensional probabilistic site-exchange cellular automaton. *Phys Rev E* 48:3168–3171
34. Kohring GA, Schreckenberg M (1992) The Domany–Kinzel cellular automaton revisited. *J Phys I (France)* 2:2033–2037
35. Lübeck S (2004) Universal scaling of non-equilibrium phase transitions. *Int J Mod Phys B* 18:3977–4118
36. Essam J (1989) Directed compact percolation: cluster size and hyperscaling. *J Phys A: Math Gen* 22:4927–4937
37. Dickman R, Tretyakov A (1995) Hyperscaling in the Domany–Kinzel cellular automaton. *Phys Rev E* 52:3218–3220
38. Bohr T, van Hecke M, Mikkelsen R, Ipsen M (2001) Breakdown of universality in transitions to spatiotemporal chaos. *Phys Rev Lett* 24:5482–5485
39. Katori M, Konno N, Tanemura H (2002) Limit theorems for the nonattractive Domany–Kinzel model. *Ann Probab* 30:933–947
40. Biham O, Middleton A, Levine D (1992) Self-organization and a dynamical transition in traffic flow models. *Phys Rev A* 46:R6124–R6127
41. Nagatani T (1993) Jamming transition in the traffic-flow model with two-level crossings. *Phys Rev E* 48:3290–3294
42. Nagatani T (1993) Effect of traffic accident on jamming transition in traffic-flow model. *J Phys A: Math Gen* 26:L1015–L1020
43. Ishibashi Y, Fukui M (1996) Phase diagram for the traffic model of two one-dimensional roads with a crossing. *J Phys Soc Jpn* 65:2793–2795
44. Boccara N, Fukš H, Zeng Q (1997) Car accidents and number of stopped cars due to road blockade on a one-lane highway. *J Phys A: Math Gen* 30:3329–3332
45. Sakakibara T, Honda Y, Horiguchi T (2000) Effect of obstacles on formation of traffic jam in a two-dimensional traffic network. *Phys A: Stat Mech Appl* 276:316–337
46. Schadschneider A, Chowdhury D, Brockfeld E, Klauck K, Santen L, Zittartz J (2000) A new cellular automata model for city traffic. In: Helbing D, Herrmann H, Schreckenberg M, Wolf DE (eds) *Traffic and Granular Flow 1999: Social, Traffic, and Granular Dynamics*. Springer, Berlin
47. Fukš H, Boccara N (1998) Generalized deterministic traffic rules. *Int J Mod Phys C* 9:1–12
48. Fukš H (1999) Exact results for deterministic cellular automata traffic models. *Phys Rev E* 60:197–202
49. Fukui M, Ishibashi Y (1999) Self-organized phase transitions in cellular automaton models for pedestrians. *J Phys Soc Jpn* 68:2861–2863
50. Fukui M, Ishibashi Y (1999) Jamming transition in cellular automaton models for pedestrians on passageway. *J Phys Soc Jpn* 68:3738–3739
51. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199–329
52. Rickert M, Nagel K (2001) Traffic simulation: Dynamic traffic assignment on parallel computers in TRANSIMS. *Future Gener Comput Syst* 17:637–648
53. Refer to the TRANSIMS Web site:  
<http://transims.tsasa.lanl.gov/>

## Books and Reviews

Here are a few articles on different problems of phase transitions in cellular automata that might be of interest to readers wishing to go in more details.

- Bagnoli F, Franci F, Rechtman R (2002) Opinion formation and phase transitions in a probabilistic cellular automaton with two absorbing states. *Lecture Notes in Computer Science*, vol 2493. Springer, pp 249–258
- Behera L, Schweitzer F (2003) On spatial consensus formation: Is the Sznajd model different from a voter model? *Int J Mod Phys C* 14:1331–1354
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199–329
- Hołyst JA, Kacperski K, Schweitzer F (2000) Phase transitions in social impact models of opinion formation. *Physica A* 285: 199–210
- Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A* 35:9971–10013
- Maerivoet S, De Moor B (2007) Non-concave fundamental diagrams and phase transitions in a stochastic traffic cellular automaton. *Eur Phys J* 42:131–140
- Takeuchi K (2006) Can the Ising critical behavior survive in non-equilibrium synchronous cellular automata? *Physica D* 223:146–150
- van Wijland F (2002) Universality class of nonequilibrium phase transition with infinitely many absorbing states. *Phys Rev Lett* 89:190 602

## Phase Transitions on Fractals and Networks

DIETRICH STAUFFER

Institute for Theoretical Physics, Cologne University,  
Köln, Germany

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Ising Model  
Fractals  
Diffusion on Fractals  
Ising Model on Fractals  
Networks  
Future Directions  
Bibliography

### Glossary

**Cluster** Clusters are sets of occupied neighboring sites.

**Critical exponent** At a critical point or second-order phase transition, many quantities diverge or vanish with a power law of the distance from this critical point; the critical exponent is the exponent for this power law.

**Diffusion** A random walker decides at each time step randomly in which direction to proceed. The resulting mean square distance normally is linear in time.

**Fractals** Fractals have a mass varying with some power of their linear dimension. The exponent of this power law is called the fractal dimension and is smaller than the dimension of the space.

**Ising model** Each site carries a magnetic dipole which points up or down; neighboring dipoles “want” to be parallel.

**Percolation** Each site of a large lattice is randomly occupied or empty.

### Definition of the Subject

At a phase transition, as a function of a continuously varying parameter (like the temperature), a sharp singularity happens in infinitely large systems, where quantities (e. g. the density) jump, vanish, or diverge.

### Introduction

Some phase transitions, like the ferromagnetic Curie point where the spontaneous magnetization vanishes, happen

in solids, and experiments often try to grow crystals very carefully such that the solid in which the transition will be observed is periodic with very few lattice faults. Other phase transitions like the boiling of water, or the liquid-vapor critical point where the density difference between a liquid and its vapor vanishes, happen in a continuum without any underlying lattice structure. Nevertheless, the critical exponents of the Ising model on a simple-cubic lattice agree well with those of liquid-vapor experiments. Impurities, which are either fixed (“quenched dilution”) or mobile (“annealed dilution”), are known to change these exponents slightly, e. g. by a factor  $1 - \alpha$ , if the specific heat diverges in the undiluted case at the critical point, i. e. if the specific heat exponent  $\alpha$  is positive. In this review we deal neither with regular lattices nor with continuous geometry, but with phase transitions on fractal and other networks. We will compare these results with the corresponding phase transitions on infinite periodic lattices like the Ising model.

### Ising Model

Ernst Ising in 1925 (then pronounced EEsing, not EYE-sing) published a model which is, besides percolation, one of the simplest models for phase transitions. Each site  $i$  is occupied by a variable  $S_i = \pm 1$  which physicists often call a spin but which may also be interpreted as a trading activity [10] on stock markets, as a “race” or other ethnic group in the formation of city ghettos [26], as the type of molecule in binary fluid mixtures like isobutyric acid and water, as occupied or empty in a lattice-gas model of liquid-vapor critical points, as an opinion for or against the government [29], ► [Opinion Dynamics and Sociophysics](#), or whatever binary choice you have in mind. Also models with more than two choices, like  $S_i = -1, 0$  and  $1$  have been investigated both for atomic spins as well as for races, opinions, . . . Two spins  $i$  and  $k$  interact with each other by an energy  $-JS_iS_k$  which is  $-J$  if both spins are the same and  $+J$  if they are the opposite of each other. Thus  $2J$  is the energy to break one bond, i. e. to transform a pair of equal spins to a pair of opposite spins. The total interaction energy is thus

$$E = -J \sum_{\langle i,k \rangle} S_i S_k, \quad (1a)$$

with a sum over all neighbor pairs. If you want to impress your audience, you call this energy a Hamiltonian or Hamilton operator, even though most Ising model publications ignore the difficulties of quantum mechanics ex-

cept for assuming the discrete nature of the  $S_i$ . (If instead of these discrete one-dimensional values you want to look at vectors rotating in two- or three-dimensional space, you should investigate the XY or Heisenberg models instead of the Ising model.)

Different configurations in thermal equilibrium at absolute temperature  $T$  appear with a Boltzmann probability proportional to  $\exp(-E/k_B T)$ , and the Metropolis algorithm of 1953 for Monte Carlo computer simulations flips a spin with probability  $\exp(-\Delta E/k_B T)$ , where  $k_B$  is Boltzmann's constant and  $\Delta E = E_{\text{after}} - E_{\text{before}}$  the energy difference caused by this flip. If one starts with a random distribution of half the spins up and half down, using this algorithm at positive but low temperatures, one sees growing domains [28]. Within each domain, most of the spins are parallel, and thus a computer printout shows large black domains coexisting with large white domains. Finally, one domain covers the whole lattice, and the other spin orientation is restricted to small clusters or isolated single spins within that domain. This self-organization (biologist may call it "emergence") of domains and of phase separation appears only for positive temperatures below the critical temperature  $T_c$  and only in more than one dimension. For  $T > T_c$  (or at all positive temperatures in one dimension) we see only finite domains which no longer engulf the whole lattice. This phase transition between long-range order below and short-range order above  $T_c$  is called the Curie or critical point; we have  $J/k_B T_c = \frac{1}{2} \ln(1 + \sqrt{2})$  on the square lattice and 0.221655 on the simple cubic lattice with interactions to the  $z$  nearest lattice neighbors;  $z = 4$  and 6, respectively. The mean field approximation becomes valid for large  $z$  and gives  $J/k_B T_c = 1/z$ . Near  $T = T_c$  the difference between the number of up and down spins vanishes as  $(T_c - T)^\beta$  with  $\beta = 1/8$  in two,  $\simeq 0.32$  in three, and  $1/2$  in six and more dimensions and in mean field approximation.

We may also influence the Ising spins through an external field  $h$  by adding

$$-h \sum_i S_i \quad (1b)$$

to the energy of Eq. (1a). This external field then pushes the spins to become parallel to  $h$ . Thus we no longer have emergence of order from the interactions between the spins, but imposition of order by the external field. In this simple version of the Ising model there is no sharp phase transition in the presence of this field; instead the spontaneous magnetization (fraction of up spins minus fraction

of down spins) smoothly sinks from one to zero if the temperature rises from zero to infinity.

## Fractals

Fractals obey a power law relating their mass  $M$  to their radius  $R$ :

$$M \propto R^D \quad (2)$$

where  $D$  is the fractal dimension. An exactly solved example are random walks (= polymer chains without interaction) where  $D = 2$  if the length of the walk is identified with the mass  $M$ . For self-avoiding walks (= polymer chains with excluded volume interaction), the Flory approximation gives  $D = (d + 2)/3$  in  $d \leq 4$  dimensions ( $D(d \geq 4) = 2$  as for random walks), which is exact in one, two and four dimensions, and too small by only about two percent in three dimensions.

We now discuss the fractal dimension of the Ising model. In an infinite system at temperatures  $T$  close to  $T_c$ , the difference  $M$  between the number of up and down spins varies as  $(T_c - T)^\beta$  while the correlation length  $\xi$  varies as  $|T - T_c|^{-\nu}$ . Thus,  $M \propto \xi^{-\beta/\nu}$ . The proportionality factor varies as the system size  $L^d$  in  $d$  dimensions since all spins are equivalent. In a finite system right at the critical temperature  $T_c$  we replace  $\xi$  by  $L$  and thus have  $M \propto L^{d-\beta/\nu} = L^D$  with the fractal dimension

$$D = d - \beta/\nu \quad (d \leq 4). \quad (3a)$$

Warning: one should not apply these concepts to spin clusters if clusters are simply defined as sets of neighboring parallel spins; to be fractals at  $T = T_c$  the clusters have to be sets of neighboring parallel spins connected by active bonds, where bonds are active with probability  $1 - \exp(-2J/k_B T)$ . Then the largest cluster at  $T = T_c$  is a fractal with this above fractal dimension.

This warning is superfluous for percolation theory (see separate reviews in this encyclopedia) where each lattice site is occupied randomly with probability  $p$  and clusters are defined as sets of neighboring occupied sites. For  $p > p_c$  one has an infinite cluster spanning from one side of the sample to the other; for  $p < p_c$  one has no such spanning cluster; for  $p = p_c$  one has sometimes such spanning clusters, and then the number of occupied sites in the largest or spanning cluster is

$$M \propto L^D; \quad D = d - \beta/\nu \quad (d \leq 6) \quad (3b)$$

with the critical exponents  $\beta$ ,  $\nu$  of percolation instead of Ising models.

These were probabilistic fractal examples, as opposed to deterministic ones like the Sierpinski carpets and gaskets, which approximate in their fractal dimensions the percolation problem. We will return to them in the Sect. “Ising Model on Fractals”.

Now, instead of asking how phase transitions produce fractals we ask what phase transitions can be observed on these fractals.

## Diffusion on Fractals

### Unbiased Diffusion

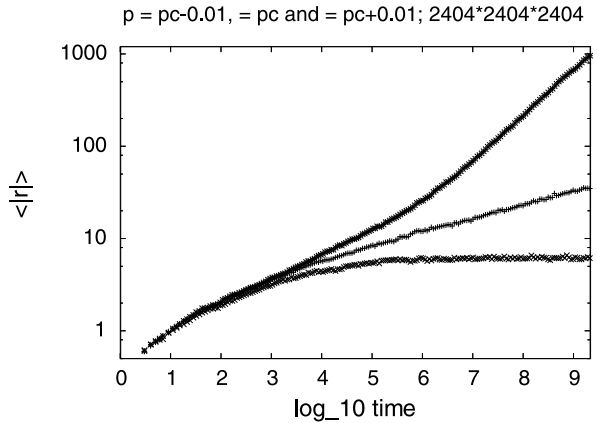
The most thoroughly investigated phase transitions on fractals are presumably random walkers on percolation clusters [17], particularly at  $p = p_c$ . This research was started by Brandt [7] but it was the later Nobel laureate de Gennes [16] who gave it the catchy name “ant in the labyrinth”. The anomalous diffusion [5,14,22] then made it famous a few years later and may have also biological applications [13].

We put an ant onto a randomly selected occupied site in the middle of a large lattice, where each site is permanently occupied (randomly with probability  $p$ ) or empty ( $1 - p$ ). At each time step, the ant selects randomly a neighbor direction and moves one lattice unit in this direction if and only if that neighbor site is occupied. We measure the mean distance

$$R(t) = \langle r(t)^2 \rangle^{1/2} \quad \text{or} \quad = \langle r(t) \rangle, \quad (4)$$

where  $\mathbf{r}$  is the vector from the starting point of the walk and the present position, and  $r = |\mathbf{r}|$  its length. The average  $\langle \dots \rangle$  goes over many such walking ants and disordered lattices. These ants are blind, that means they do not see from their old place whether or not the selected neighbor site is accessible (occupied) or prohibited (empty). (Also myopic ants were grown which select randomly always an occupied neighbor since they can see over a distance of one lattice unit.) The squared distance  $r^2$  is measured by counting how often the ant moved to the left, to the right, to the top, to the bottom, to the front, or to the back on a simple cubic lattice.

The problem is simple enough to be given to students as a programming project. They then should find out by their simulations that for  $p < p_c$  the above  $R$  remains finite while for  $p > p_c$  it goes to infinity as  $\sqrt{t}$ , for sufficiently long times  $t$ . But even for  $p > p_c$  it may happen that for a single ant the distance remains finite: If the starting point happened to fall on a finite cluster, then  $R(t \rightarrow \infty)$  measures the radius of that cluster. Let  $\mu$  be the



Phase Transitions on Fractals and Networks, Figure 1

Log-log plot for unbiased diffusion at (middle curve), above (upper data) and below (lower data) the percolation threshold  $p_c$ . We see the phase transition from limited growth at  $p_c - 0.01$  to diffusion at  $p_c + 0.01$ , separated by anomalous diffusion at  $p_c$ . Average over 80 lattices with 10 walks each

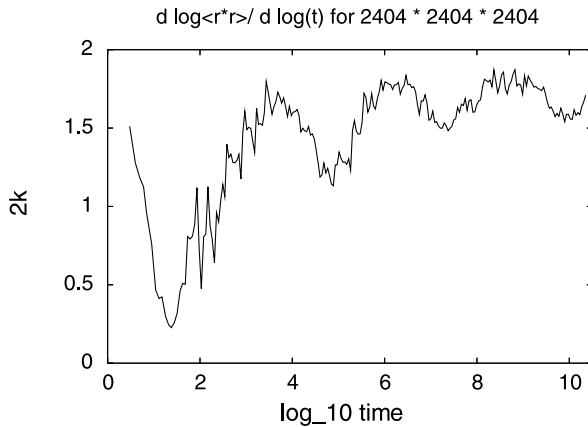
exponent for the conductivity if percolation is interpreted as a mixture of electrically conducting and insulating sites. Then right at  $p = p_c$ , instead of a constant or a square-root law, we have anomalous diffusion:

$$R \propto t^k, \quad k = (\nu - \beta/2)/(2\nu + \mu - \beta), \quad (5)$$

for sufficiently long times. This exponent  $k$  is close to but not exactly  $1/3$  in two and  $1/5$  in three dimensions.  $\beta$  and  $\nu$  are the already mentioned percolation exponents. If we always start the ant walk on the largest cluster at  $p = p_c$  instead of on any cluster, the formula for the exponent  $k$  simplifies to  $\nu/(2\nu + \mu - \beta)$ . The theory is explained in detail in the standard books and reviews [8,17]. We see here how the percolative phase transition influences the random walk and introduces there a transition between diffusion for  $p > p_c$  and finite motion for  $p < p_c$ , with the intermediate “anomalous” diffusion (exponent below  $1/2$ ) at  $p = p_c$ . Figure 1 shows this transition on a large cubic lattice.

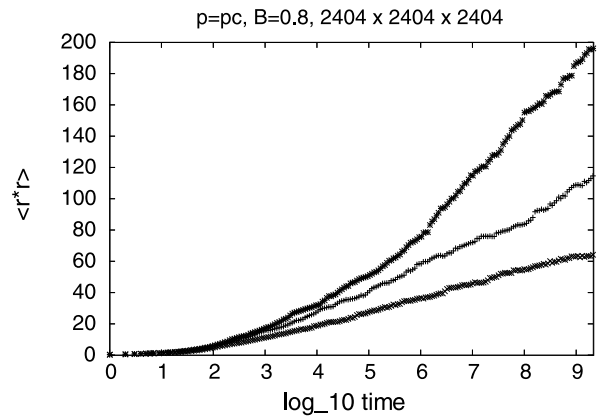
### Biased Diffusion

Another type of transition is seen in biased diffusion, also for  $p > p_c$ . Instead of selecting all neighbors randomly, we do that only with probability  $1 - B$ , while with probability  $B$  the ant tries to move in the positive  $x$ -direction. One may think of an electron moving through a disordered lattice in an external electric field. For a long time experts discussed whether for  $p > p_c$  one has a drift behavior (dis-



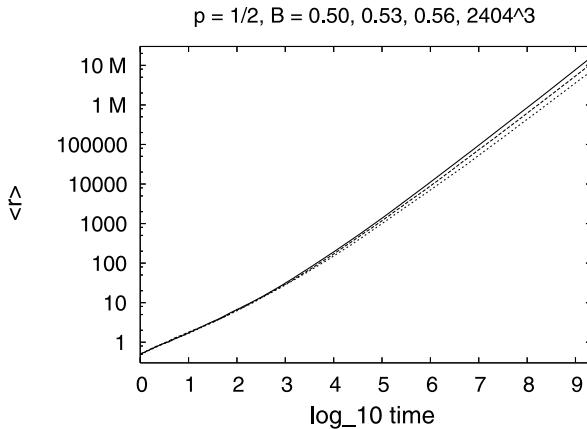
**Phase Transitions on Fractals and Networks, Figure 2**

Log-periodic oscillation in the effective exponent  $k$  for biased diffusion;  $p = 0.725$ ,  $B = 0.98$ . The limit  $k = 1$  corresponds to drift. 80 lattices with 10 walks each



**Phase Transitions on Fractals and Networks, Figure 4**

Biased diffusion at  $p = p_c$  (middle curve) and  $p = p_c \pm 0.01$  (upper and lower data) for bias  $B = 0.8$ ; 80 lattices with 10 walks each



**Phase Transitions on Fractals and Networks, Figure 3**

Difficulties at transition from drift (small bias, upper data) to slower motion (large bias, lower data); 80 lattices with 10 walks each

tance proportional to time) for small  $B$ , and a slower motion for larger  $B$ , with a sharp transition at some  $p$ -dependent  $B_c$ . In the drift regime one may see log-periodic oscillations  $\propto \sin(\text{const} \log t)$  in the approach towards the long-time limit, Fig. 2. Such oscillations have been predicted for stock markets [19], where they could have made us rich, but for diffusion they hamper the analysis. They come presumably from sections of occupied sites which allow motion in the biased direction and then end in prohibited sites [20].

Even in a region without such oscillations, Fig. 3 shows no clear transition from drift to no drift; that transition could only be seen by a more sophisticated analysis which

showed for the  $p$  of Fig. 3 that the reciprocal velocity, plotted vs.  $\log(\text{time})$ , switches from concave to convex shape at  $B_c \simeq 0.53$ . Fortunately, only a few years after these simulations [11] the transition was shown to exist mathematically [6].

These simulations were made for  $p > p_c$ ; at  $p = p_c$  with a fractal largest cluster, drift seems impossible, and for a fixed  $B$  the distance varies logarithmically, with a stronger increase slightly above  $p_c$  and a limited distance slightly below  $p_c$ , Fig. 4.

### Ising Model on Fractals

What happens if we set Ising spins onto the sites of a fractal? In particular, but also more generally, what happens to Ising spins on the occupied site of a percolation lattice, when each site is randomly occupied with probability  $p$ ? In this case one expects three sets of critical exponents describing how the various quantities diverge or vanish at the Curie temperature  $T_c(p)$ . For  $p = 1$  one has the standard Ising model with the standard exponents. If  $p_c$  is the percolation threshold where an infinite cluster of occupied sites starts to exist, then one has a second set of exponents for  $p_c < p < 1$ , where  $0 < T_c(p) < T_c(p = 1)$ . Finally, for zero temperature as a function of  $p - p_c$  one has the percolation exponents as a third set of critical exponents. (If  $p = p_c$  and the temperature approaches  $T_c(p_c) = 0$  from above, then instead of powers of  $T - T_c$  exponential behavior is expected.) In computer simulations, the second set of critical exponents is difficult to observe; due to limited accuracy the effective exponents have a tendency to vary continuously with  $p$ .



The behavior at zero temperature is in principle trivial: each cluster of occupied neighbors has parallel spins, the spontaneous magnetization is given by the largest cluster while the many finite clusters cancel each other in their magnetization. However, the existence of several infinite clusters at  $p = p_c$  disturbs this argument there; presumably the total magnetization (i. e. not normalized at magnetization per spin) is a fractal with the fractal dimension of percolation theory.

Deterministic fractals, instead of the random “incipient infinite cluster” at the percolation threshold, may have a positive  $T_c$  and then allow a more usual study of critical exponents at that phase transition. Koch curves and Sierpinski structures have been intensely studied in that aspect since decades. To build a Sierpinski carpet we take a square, divide each side into three thirds such that the whole square is divided into nine smaller squares, and then we take away the central square. On each of the remaining eight smaller squares this procedure is repeated, dividing each into nine squarelets and omitting the central squarelet. This procedure is repeated again and again, mathematically ad infinitum. Physicists like more to think in terms of a fixed distance and would rather imagine each square to be enlarged in each direction by a factor three with the central square omitted; and then again and again this enlargement is repeated. In this way we grow a large structure built by squares of unit area.

Unfortunately, the phase transitions on these fractals depend on details and are not already fixed if the fractal dimension is fixed. Also other properties of the fractals like their “ramification” are important [15]; see [3] for recent work. This is highly regrettable since modern statistical physics is not restricted to three dimensions. Models were studied also in seven and in minus two dimensions, in the limits of dimensionality going to infinity or to zero, or for non-integral dimensionality. (Similarly, numbers were generalized from positive counts to negative integers, to rational and irrational numbers, and finally to imaginary/complex numbers.) It would have been nice if these fractals would have been models for these non-integral dimensions, giving one and the same set of critical exponents once their fractal dimension is known. Regrettably, we had to give up that hope.

Many other phase transitions, like those of Potts or voter models, were studied on such deterministic fractals, but are not reviewed here. We mention that also percolation transitions exist on Sierpinski structures [24]. Also, various hierarchical lattices different from the above fractals show phase transitions, if Ising spins are put on them; the reader is referred to [18,25] for more litera-

ture. As to our knowledge most recent example we mention [21] that Ising spins were also thrown into the sandpiles of Per Bak, which show self-organized criticality.

## Networks

### Definitions

While fractals were a big physics fashion in the 1980s, networks are now a major physics research field. Solid state physics requires nice single crystals where all atoms sit on a periodic lattice. In fluids they are ordered only over shorter distances but still their forces are restricted to their neighbors. Human beings, on the other hand, form a regular lattice only rarely, e. g. in a fully occupied lecture hall. In a large crowd they behave more like a fluid. But normally each human being may have contacts with the people in neighboring residences, with other neighbors at the work place, but also via phone or internet with people outside the range of the human voice. Thus social interactions between people should not be restricted to lattices, but should allow for more complex networks of connections.

One may call Flory’s percolation theory of 1941 a network, and the later random graphs of Erdős and Rényi (where everybody is connected with everybody, albeit with a low probability) belong to the same “universality class” (same critical exponents) as Flory’s percolation. In Kauffman’s random Boolean network of 1969, everybody has  $K$  neighbors selected randomly from the  $N$  participants. Here we concentrate on two more recent network types, the small-world [31] and the scale-free [1] networks of 1998 and 1999 respectively (with a precursor paper of economics Nobel laureate Simon [27] from 1955).

The small-world or Watts–Strogatz networks start from a regular lattice, often only a one-dimensional chain. Then each connection of one lattice site to one of its nearest neighbors is replaced randomly, with probability  $p$ , by a connection to a randomly selected other site anywhere in the lattice. Thus the limits  $p = 0$  and  $1$  correspond to regular lattices and roughly random graphs, respectively.

In this way everybody may have exactly two types of connections, to nearest neighbors and to arbitrarily far away people. This unrealistic feature of small-world networks is avoided by the scale-free networks of Barabási and Albert [1], defined only through topology with (normally) no geometry involved:

We start with a small set of fully connected people. Then more people join the network, one after the other. Each new member selects connections to exactly  $m$  already

existing members of the network. These connections are not random but follow preferential attachment: The more people have selected a person to be connected with in the past, the higher is the probability that this person is selected by the newcomer: the rich get richer, famous people attract more followers than normal people. In the standard Barabási–Albert network, this probability is proportional to the number of people who have selected that person. In this case, the average number of people who have been selected by  $k$  later added members varies as  $1/k^3$ . A computer program is given e. g. in [28].

These networks can be undirected (the more widespread version) or directed (used less often.) For the undirected or symmetric case, the connections are like friendships: If A selects B as a friend, then B also has A as a friend. For directed networks, on the other hand, if A has selected B as a boss, then B does not have A as a boss, and the connection is like a one-way street. Up to  $10^8$  nodes were simulated in scale-free networks. We will now check for phase transitions on both directed and undirected networks.

### Phase Transitions

The Ising model in one dimension does not have a phase transition at some positive critical temperature  $T_c$ . However, its small-world generalization, i. e. the replacement of small fraction of neighbor bonds by long-range bonds, produces a positive  $T_c$  with a spontaneous magnetization proportional to  $(T_c - T)^\beta$ , and a  $\beta$  smaller than the 1/8 of two dimensional lattices [4].

The Solomon network is a variant of the small-world network: Each person has one neighborhood corresponding to the workplace and another neighborhood corresponding to the home [23]. It was suggested and simulated by physicists Solomon and Malarz, respectively, before Edmonds [12] criticized physicists for not having enough “models which explicitly include actions and effects within a physical space as well as communication and action within a social space”. Even in one dimension a spontaneous magnetization was found.

On Barabási–Albert (scale-free) networks, Ising models were found [2] for small  $m$  and millions of spins to have a spontaneous magnetization for temperatures below some critical temperature  $T_c$  which increases logarithmically with the number  $N$  of spins:  $k_B T_c/J \simeq 2.6 \ln(N)$  for  $m = 5$ .

Here we had undirected networks with symmetric couplings between spins:  $\text{actio} = -\text{reactio}$ , as required by Newton. Ising spins on directed networks, on the other hand, have no well-defined total energy, though each sin-

gle spin may be influenced as usual by its  $m$  neighbor spins. If in an isolated pair of spins  $i$  and  $k$  we have a directed interaction in the sense that spin  $k$  tries to align spin  $i$  into the direction of spin  $k$ , while  $i$  has no influence on  $k$ , then we have a perpetual mobile: Starting with the two spins antiparallel, we first flip  $i$  into the direction of  $k$ , which gives us an energy  $2J$ . Then we flip spin  $k$  which does not change the energy. Then we repeat again and again these two spin flips, and gain an energy  $2J$  for each pair of flips: too nice to be true. The violations of Newton’s symmetry requirement makes this directed network applicable to social interactions between humans, but not to forces between particles in physics.

On this directed Barabási–Albert network, the ferromagnetic Ising spins gave no spontaneous magnetization, but the time after which the magnetization becomes zero (starting from unity) becomes very long at low temperatures, following an Arrhenius law [30]: time proportional to  $\exp(\text{const}/T)$ . Also on a directed lattice such Arrhenius behavior was seen while for directed random graphs and for directed small-world lattices a spontaneous magnetization was found [30]. A theoretical understanding for these directed cases is largely lacking.

Better understood is the percolative phase transition on scale-free networks (see end of Sect. “Introduction” for definition of percolation). If a fraction  $1 - p$  of the connections in an undirected Barabási–Albert network is cut randomly, does the remaining fraction  $p$  keep most of the network together? It does, for large enough networks [9], since the percolation threshold  $p_c$  below which no large connected cluster survives, goes to zero as  $1/\log(N)$  where  $N$  counts the number of nodes in the network. This explains why inspite of the unreliability of computer connections, the internet still allows most computers to reach most other computers in the word: If one link is broken, some other link may help even though it may be slower [1]. (For intentional [9] cuts in hierarchical networks one may have a finite percolation threshold [32].)

### Future Directions

We reviewed here a few phase transitions, and ignored many others. At present most interesting for future research seem to be the directed networks, since they have been investigated with methods from computational physics even though they are not part of usual physics, not having a global energy. A theoretical (i. e. not numerical) understanding would help.

We thank K. Kulakowski for comments.

## Bibliography

1. Albert R, Barabási AL (2002) Rev Mod Phys 74:47; Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Phys Repts 424:175
2. Aleksiejuk A, Holyst JA, Stauffer D (2006) Physica A 310:260; Dorogovtsev SN, Goltsev AV, Mendes JFF (2002) Phys Rev E 66:016104. arXiv:0705.0010 [cond-mat.stat-mech]; Bianconi G (2002) Phys Lett A 303:166
3. Bab MA, Fabricius G, Albano EV (2005) Phys Rev E 71:036139
4. Barrat A, Weigt M (2000) Eur Phys J B 13:547; Gitterman M (2000) J Phys A 33:8373; Pękalski A (2001) Phys Rev E 64:057104
5. Ben-Avraham D, Havlin S (1982) J Phys A 15:L691
6. Berger N, Ganten N, Peres Y (2003) Probab Theory Relat Fields 126:221
7. Brandt WW (1975) J Chem Phys 63:5162
8. Bunde A, Havlin S (1996) Fractals and Disordered Systems. Springer, Berlin
9. Cohen R, Erez K, ben-Avraham D, Havlin S (2000) Phys Rev Lett 85:4626; Cohen R, Erez K, ben-Avraham D, Havlin S (2001) Phys Rev Lett 86:3682; Callaway DS, Newman MEJ, Strogatz SH, Watts DJ (2000) Phys Rev Lett 85:5468
10. Cont R, Bouchaud J-P (2000) Macroecon Dyn 4:170
11. Dhar D, Stauffer D (1998) Int J Mod Phys C 9:349
12. Edmonds B (2006) In: Billari FC, Fent T, Prsakwetz A, Scheffran J (eds) Agent-based computational modelling. Physica-Verlag, Heidelberg, p 195
13. Frey E, Kroy K (2005) Ann Physik 14:20
14. Gefen Y, Aharony A, Alexander S (1983) Phys Rev Lett 50:77
15. Gefen Y, Mandelbrot BB, Aharony A (1980) Phys Rev Lett 45:855
16. de Gennes PG (1976) Rech 7:916
17. Havlin S, Ben Avraham D (1987) Adv Phys 36:395; Havlin S, Ben Avraham D (2002) Adv Phys 51:187
18. Hinczewski M, Berker AN (2006) Phys Rev E 73:066126
19. Johansen A, Sornette D (1999) Int J Mod Phys C 10:563
20. Kirsch A (1999) Int J Mod Phys C 10:753
21. Koza Z, Ausloos M (2007) Physica A 375:199
22. Kutner R, Kehr K (1983) Phil Mag A 48:199
23. Malarz K (2003) Int J Mod Phys C 14:561
24. Monceau P, Hsiao PY (2004) Phys Lett A 332:310
25. Rozenfeld HD, Ben-Avraham D (2007) Phys Rev E 75:061102
26. Schelling TC (1971) J Math Sociol 1:143
27. Simon HA (1955) Biometrika 42:425
28. Stauffer D, Moss de Oliveira S, de Oliveira PMC, Sá Martins JS (2006) Biology, Sociology, Geology by Computational Physicists. Elsevier, Amsterdam
29. Sznajd-Weron K, Sznajd J (2000) Int J Mod Phys C 11:1157
30. Sánchez AD, López JM, Rodríguez MA (2002) Phys Rev Lett 88:048701; Sumour MA, Shabat MM (2005) Int J Mod Phys C 16:585; Sumour MA, Shabat MM, Stauffer D (2006) Islamic Univ J (Gaza) 14:209; Lima FWS, Stauffer D (2006) Physica A 359:423; Sumour MA, El-Astal AH, Lima FWS, Shabat MM, Khalil HM (2007) Int J Mod Phys C 18:53; Lima FWS (2007) Comm Comput Phys 2:522 and (2008) Physica A 387:1545; 3503
31. Watts DJ, Strogatz SH (1998) Nature 393:440
32. Zhang Z-Z, Zhou S-G, Zou T (2007) Eur Phys J B 56:259

## Philosophy of Science, Mathematical Models in

ZOLTAN DOMOTOR

University of Pennsylvania, Philadelphia, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Mathematical Models: What Are They?](#)

[Philosophical and Mathematical Structuralism](#)

[Three Approaches to Applying Mathematical Models](#)

[Validating Mathematical Models](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Philosophy of science** Broadly understood, *philosophy of science* is a branch of philosophy that studies and reflects on the presuppositions, concepts, theories, arguments, methods and aims of science. Philosophers of science are concerned with general questions which include the following: What is a scientific theory and when can it be said to be confirmed by its predictions? What are mathematical models and how are they validated? In virtue of what are mathematical models representations of the structure and behavior of their target systems? In sum, a major task of philosophy of science is to analyze and make explicit common patterns that are implicit in scientific practice.

**Mathematical model** Stated loosely, *models* are simplified, idealized and approximate representations of the structure, mechanism and behavior of real-world systems. From the standpoint of set-theoretic model theory, a *mathematical model* of a target system is specified by a nonempty set – called the model's *domain*, endowed with some operations and relations, delineated by suitable *axioms* and intended empirical *interpretation*. No doubt, this is the simplest definition of a model that, unfortunately, plays a limited role in scientific applications of mathematics. Because applications exhibit a need for a large variety of vastly different mathematical structures – some topological or smooth, some algebraic, order-theoretic or combinatorial, some measure-theoretic or analytic, and so forth, no useful overarching definition of a mathematical model is known even in the edifice of modern category theory. It is difficult to come up with a workable

concept of a mathematical model that is adequate in most fields of applied mathematics and anticipates future extensions.

**Target system** There are many definitions of the concept of ‘system’. Here by a *target system* we mean an effectively isolated (physical, biological, or other empirical) part of the universe – made to function or run by some internal or external causes, whose interactions with the universe are strictly delineated by a fixed (input and output) interface, and whose structure, mechanism, or behavior are the objects of mathematical modeling. Changes produced in the target system are presumed to be externally detectable via measurements of the system’s characterizing quantitative properties.

### Definition of the Subject

Models are indispensable scientific tools in generating information about the world. Concretely, scientists rely on models for explanation and prediction. Recent years have seen intensive attempts to extend the art of modeling and simulation to a vast range of application areas. Fostered by inexpensive computers and software, and the accelerating growth of mathematical knowledge, in contemporary applied research reference to models is far more frequent than reference to theories or principles. In applications, the term ‘model’ is used in a wide variety of senses, including mathematical, physical, mental, and computational models. Our focus here is on mathematical models, their structure, representational role, and validation.

Currently, many philosophers of science are interested in two major aspects of mathematical models: their *nature* and representational *role*. The question “What are mathematical models?” is answered in two fundamentally different but closely related ways:

1. In a seemingly natural way, by viewing mathematical models as families of *equations* of some kind, accompanied with certain empirical interpretations that link them to their target systems. It turns out that this simple conception, called the *received view* or the *syntactic* approach, presents several troubling representational and interpretational problems.
2. The so-called *structuralist* or *semantic* answer takes mathematical models of target systems to be suitable set-theoretic structures or generally objects in a specific dynamical (or other) category. In order to pursue this approach, model builders and users need to be able to understand how complex notions, relevant

to modeling, are defined in terms of the model’s structure.

One of the most useful general results on the nature of mathematical models is the following. If a given semantic formulation of a model involves also a specification of the *solution space* of some equations, then in this case the equational and structural conceptions of models are formally equivalent. Indeed, equations uniquely characterize their solution spaces and these in turn determine the associated semantic model. Conversely, if the semantic model specifies an abstract solution space, then the latter delineates a system of characterizing equations. This type of *Galois correspondence* between equations and their solutions has been established in many linear and even in some non-linear settings.

There is a major counter to this syntactic-semantic dilemma. Structuralists are quick to point out that the mathematical models of exchange economy,  $n$ -person games, probability and decision, and so forth, are *autonomous* set-theoretic structures that are not associated with any system of equations. For example, recall that a classical probability model of a statistical experiment is defined by a triple, consisting of a set together with a designated field of its subsets – forming an underlying measurable space, and a probability measure thereon. Remarkably, even though in this case there are no equations to consider, thanks to the powerful *Stone–Gelfand duality* result, every probability model has an associated algebraic counterpart model, given by a linear space (thought of as the space of bounded random variables on the model’s underlying measurable space) and a positive linear functional on it (viewed as the expectation functional induced by the measure). Statisticians in particular (e.g., [21]) prefer to build their models of experiments in a ‘dual’, computationally stronger, algebraic setting. It happens quite often that autonomous mathematical models arise in ‘adjoint’ or paired geometric/algebraic formulations. Characteristically, models of this nature tend to form impressively versatile ‘mathematical universes’ for all of the mathematics that model users may need in a given area of application.

The question “What is the role of mathematical models in scientific practice?” is answered by describing how models are conceived, constructed, and used in various applications. Since models are mathematical constructs, in addition to being objects of a formal inquiry, they are also involved in epistemic relations, expressing intended uses.

Models do not and need not match reality in all of its aspects and details to be adequate. A mathematical

model is usually developed for a specific class of target systems, and its validity is determined relative to its intended applications. A model is considered *valid* within its intended domain of applicability provided that its predictions in that domain fall within an acceptable range of error, specified prior to the model's development or identification.

To construct a mathematical model of a target system, it is typically necessary to formalize the system's decisive causal mechanisms and behavior with special regards to the model's *structural stability* and *mutability* into a larger network of models that includes not only additional cause-effect relations but also a broad range of deterministic and stochastic perturbations. In the next section we consider a simple class of dynamical models in physics that is also relevant to modeling in other disciplines.

## Introduction

The subject of mature mathematical models in the form of *equations* has its roots in post-Newtonian developments of classical mechanics, hydrodynamics, electromagnetism, and kinetic theory of gases. It came on the scene of applied mathematics gradually, during the *analytic period* before 1880, thanks to the innovative efforts of great scientists, including, among many others, the Swiss mathematician Leonhard Euler (1707–1783), Italian–French mathematician Louis Joseph Lagrange (1736–1813), French astronomer-physicist Pierre Simon de Laplace (1749–1827), Scottish physicist James Clerk Maxwell (1831–1879), English physicist Lord John William Strutt Rayleigh (1842–1919), and the Austrian physicist Ludwig Boltzmann (1844–1906). It was the genius of the French mathematician Henri Poincaré (1854–1912) that generated many of our current topological and differential methods of mathematical modeling in the world of dynamical systems. Over the past 100 years or so, mathematical models have evolved to become the basic tools in a wide variety of disciplines, including not only most of physical sciences, but also chemical kinetics, population dynamics, economics, sociology, and psychology.

The many different theoretical areas of natural and social sciences have led to the development of a large assortment of mathematical models, including but not limited to descriptive vs. normative, static vs. dynamic, phenomenological vs. process-based, discrete vs. continuous, deterministic vs. stochastic, linear vs. nonlinear, finite- vs. infinite-dimensional, difference vs. differential, and topological vs. measure-theoretic models. Using category theory, efforts have been made to construct general theories of

mathematical models of which models of logical systems and dynamical systems are special cases.

There has been a renewed interest also among philosophers of science in the problems of structure and function of abstract models. (See, for example, [3,5,10,13,25], and [33].) Prime questions about abstract models good many philosophers ask and attempt to answer include the following:

- (i) *Ontology of models*: What, precisely, are mathematical models and how are they used in science? Are they structures in the sense of classical set theory, modern category theory, or something else, belonging, e.g., to the nebulous world of fictional entities or human constructs?
- (ii) *Semantics of models*: In virtue of what are mathematical models representations of the structure, causal mechanisms and behavioral regimes of their target systems? Is it in virtue of some (possibly partial) 'isomorphisms' holding between mathematical and physical domains or by reason of certain designated 'similarities', analogies, or resemblance relations between models and aspects of the world, or because of 'empirical interpretations' of (parts of) the representing model's mathematical vocabulary – implying quantitative claims about the world that can be corroborated by empirical data, or in virtue of yet something else, such as homology or physical instantiation? Since models appear to be the main vehicles in the pursuit of scientific knowledge, philosophers are also interested in analyzing the truth conditions of semantic relations of a more general nature, such as "Researcher  $\mathcal{R}$  uses model  $\mathcal{M}$  to represent target system  $\mathcal{S}$  for purpose  $\Pi$ ".
- (iii) *Validation of models*: How are mathematical models validated? Is validation just a procedure in which the model's predictions are simply compared with a set of observations within the model's domain of applicability, or is it a comprehensive, all-out testing to determine the degree of agreement between the model and its target system in terms of internal structure, cause-effect relationships, and predictions?

Needless to add, these and many other questions to be examined below do not belong to science per se; they are *about* science. In this sense, philosophy of science is a second-order discipline, addressing the practices, methods and aims of the various sciences. However, it is clear from [24] that second-order questions about science are investigated also by scientists themselves.



Before considering the details of *models* of mathematical models, we begin by characterizing a particular use of the term *mathematical model* which, although not representative of the more careful formulations of the majority of philosophers of science, is nevertheless the most common practice encountered in physics, engineering and economics. Most physicists, engineers and mathematically oriented social scientists understand mathematical models to be systems of (algebraic, difference, differential, integral or other) *equations* (linking time- and spacetime-dependent quantities, and parameters), derived from first principles under various idealizing and simplifying scenarios of target systems or induced by available observational data and ‘empirical laws’.

As an example illustrating the equational conception of mathematical models, we shall consider briefly the most familiar simple classical planar *pendulum model*, describing an undamped pendulum’s dynamical behavior. (Additional examples can be found in [12].) It is specified by the autonomous, deterministic second-order nonlinear differential equation

$$\frac{d^2\theta}{dt^2} + \frac{g}{\ell} \sin \theta = 0,$$

in which the time-dependent *indeterminate*  $\theta$  represents the target pendulum’s variable angle from its downward vertical to the pivoting rod, to which a bob of mass  $m$  is attached at its swinging end. Coefficient  $g$  denotes the homogeneous gravitational acceleration acting on the pendulum’s bob downward, and coefficient  $\ell$  captures the length of the pendulum’s idealized ‘massless’ and perfectly stiff rod. These constant coefficients are needed for individuating the target pendulum in its *classical* gravitational ambience. Under the accompanying physical interpretation, fixed by the pendulum’s *idealizing scenario* (involving significant idealizations of friction, torque, resistance, and elasticity) and first-principle framework, the all-important observable quantity is the pendulum’s total energy (Hamiltonian)

$$H\left(\theta, \frac{d\theta}{dt}\right) =_{\text{df}} m\ell^2 \left( \frac{1}{2} \left( \frac{d\theta}{dt} \right)^2 - \frac{g}{\ell} \cos \theta \right).$$

The first term in the differential equation above encodes inertia and the second term stands for gravity. Recall that the pendulum’s rod is suspended from a pivot point, around which it oscillates or rotates in a vertical plane without surface resistance and forcing, so that there are no extra *additive* terms in the equation for the effects of

friction and torque. Coarsely speaking, in general a model is a simplified representation of a real-world system of interest for designated scientific purposes. Although the target system has many important features, not all of them can and should be included in the model for reasons of tractability and limited epistemic import. And those that are included, often involve drastic idealizations, known to be empirically false.

It has long been known that in general the foregoing differential equation does not have a closed-form solution in terms of traditional elementary functions. For general analytic solutions, Jacobi’s periodic elliptic functions are needed, with values knowable only with specified degrees of accuracy from mathematical tables or computations performed by special computer programs. The gap between theoretically granted solutions and their approximate variants has led Harald Atmanspacher and Hans Primas [2] to advocate a dichotomy between *states of reality* and *states of knowledge*. In a nutshell, derivation of highly theoretical equations of motion (providing maximal information) offers a so-called *ontic* (endophysical) view of modes of being of target systems, whereas (statistical) approximation and measurement procedures give an *epistemic* (exophysical) perspective on real-world systems, involving errors and updating. It is well known that the nonlinear equation above has several geometric types of solution that capture all sorts of swinging and rotating motions, and states of rest – for the most part knowable only approximately.

Moving beyond the important ontic vs. epistemic dichotomy in modeling, note that because the foregoing pendulum equation’s nonlinear component is representable by the infinite series  $\sin \theta = \theta - \theta^3/(3!) + \theta^5/(5!) - \dots$  and since for *small* deflections (less than  $5^\circ$ ) of the pendulum’s rod from the vertical the values  $\sin \theta(t)$  of the angle quantity are very nearly equal to  $\theta(t)$ , we can substitute  $\theta$  for  $\sin \theta$  in the equation and forget about the higher-order polynomial terms in the infinite series. So, at the cost of obvious approximation errors, we obtain the *basic linear pendulum* differential equation  $d^2\theta/dt^2 + g/\ell \theta = 0$  that is easy to solve. It is an elementary exercise to show that its smooth closed-form solutions are given by parametrized trigonometric position functions of the form

$$\theta(t) = \theta(0) \cos(\omega t) + \frac{\dot{\theta}(0)}{\omega} \sin(\omega t)$$

for all time instants  $t$ , with initial conditions  $\theta(0)$  and  $\dot{\theta}(0) =_{\text{df}} \frac{d\theta}{dt}(t)|_{t=0}$ , and parameter  $\omega =_{\text{df}} \sqrt{\ell : g}$ , characterizing the pendulum’s frequency of oscillation. In the study of differential equations and their solutions it is stan-

dard to adopt the so-called *flow* point of view. Simply, instead of studying a particular solution for a given initial value, the entire *solution space*

$$\left\{ \theta \in \mathbb{C}^\infty(\mathbb{T}) \mid \frac{d^2\theta}{dt^2} + \frac{g}{\ell}\theta = 0 \right\}$$

is studied, preferably in a parametrized form. Granted that the model is ‘correct’, this subspace of the space  $\mathbb{C}^\infty(\mathbb{T})$  of smooth real-valued functions on a designated continuum-time domain  $\mathbb{T}$  provides the investigator with complete information about the target pendulum’s possible angular positions and dynamical behavior. Clearly, the foregoing solution space includes many relativistically forbidden solutions that encode physically impossible behaviors – individuated by the pendulum’s superluminal velocities. The key idea here is that in the absence of complete knowledge of the model’s empirical *domain of applicability*, a researcher may be in an epistemic danger of trying to physically interpret (in terms of behavior) some of the solutions – countenanced by the model, that are actually meaningless in the physical world.

To remedy the situation, the foregoing classical simple pendulum model must be extended to the *relativistic* case (studied, e.g., in [11]), having the form

$$\frac{d^2\theta}{dt^2} + \left( 1 - \left( \frac{\ell}{c} \frac{d\theta}{dt} \right)^2 \right) \frac{g}{\ell} \sin \theta = 0,$$

where parameter  $c$  denotes the speed of light. Because formal models usually possess limited empirical domains of applicability, in comprehensive treatments of target system behavior several different, closely related models may become necessary.

Real-world systems tend to have many representing models and these models often possess a surplus content, which supports their mutations and extensions in unexpected ways. For example, in the presence of randomness or ‘noise’, a classical stochastic pendulum model provides the most appropriate representation of randomly perturbed motion. A typical model of the behavior of simple pendulums affected by ‘noise’ is presented by the stochastic differential equation

$$\frac{d^2X}{dt^2} + (1 + \varepsilon W)\omega^2 \sin X = 0,$$

where  $X$  denotes the stochastic position indeterminate,  $\varepsilon > 0$  is a parameter with small values, and  $W$  is a (e.g., Gaussian) stochastic process, capturing the pendulum’s random perturbation. Naturally, to extract information

from a stochastic pendulum model, the investigator has to calculate the moments of target pendulum’s positions or consider the transition probability density that allows to calculate the conditional probability of a future position of the pendulum’s bob, given its position at a designated starting time.

In many applications, the same basic pendulum differential equation drops out of a wholly different idealizing scenario of, say, a coupled mass-spring system, consisting of a (point) mass attached to a spring at one end, where the other end of the spring is tied to a fixed frame. Other prominent examples of systems, whose dynamical behavior is also modeled by ‘pendulum equations’, include electric circuits, chemical reactions and interacting biological populations with oscillatory behaviors. Of course, in each application, the indeterminate and individuating parameters are interpreted differently.

These examples illustrate clearly that mathematical models are characteristically generic or *protean* with respect to real-world systems, meaning that often the same mathematical model applies to different empirical situations, admits several different empirical interpretations, and is functioning both as an investigative instrument and as an object of mathematical inquiry. Although, broadly speaking, models can be representations of a particular *token* target system (canonical examples are cosmological models) or of a *stereotype* class of systems (e.g., pendulums), they are always representations of some particular structure of a phenomenon, mechanism or behavior. Needless to add, empirically meaningful deductions from the representing equations are guided not just by the free-standing equational structure of pure mathematics but also by the accompanying empirical interpretation, encapsulated in part by the target system’s idealizing scenario. Along these lines, in [23] Saunders Mac Lane asserts that “mathematics is protean science; its subject matter consists of those structures which appear (unchanged) in different scientific contexts”.

On this picture, scientists construct formal models in the form of (differential, partial differential, stochastic differential, etc.) equations, proceeding in a simple *top-down* analytic fashion – using first principles (e.g., Newton’s, Kirchhoff’s and other laws) and idealizing scenarios or ‘empirical rules’ that are not part of any ambient theory, or in a more involved *bottom-up*, *data-driven* synthetic manner, starting from experimental (e.g., time series) data, the extant stage of knowledge, analogies with other systems, and background assumptions. In a total absence of any data or prior knowledge pertaining to the target system, it is inappropriate to consider mathematical models at all. This raises several additional foundational questions:

- (i) What is at stake, if anything, in conceiving mathematical models as empirically interpreted equations of some sort?
- (ii) How is it that the causal mechanisms and behaviors of real-world systems can be so effectively studied in terms of solutions of various equations that employ highly idealized assumptions, known to be false?
- (iii) Since most (nonlinear) equations arising in science are solvable only approximately and in a discrete range of values, what is their epistemic import? Recall that predictions from these equations may have to be obtained by brute-force numerical methods that require more computational power than the scientists are likely to have. Thus, only a Laplacean superscientist with unlimited cognitive capacities and computational resources could make full use of such equations. In contrast, although a real-world scientist does start with such “true-in-heaven” equations but then he or she quickly modifies them to make them computationally easier to extract predictions from them.

We are now ready to start discussing some of the answers to these and other previously listed philosophical questions about mathematical models.

### Mathematical Models: What Are They?

We begin, as a way of entering the subject of mathematical models by clarifying the dichotomy between *equational* (syntactic) and *structural* (semantic) conceptions of models. The standard view among most theoretical physicists, engineers and economists is that mathematical models are *syntactic* (linguistic) items, identified with particular systems of equations or relational statements. From this perspective, the process of solving a designated system of (algebraic, difference, differential, stochastic, etc.) equations of the target system, and interpreting the particular solutions directly in the context of predictions and explanations are primary, while the mathematical structures of associated state and orbit spaces, and quantity algebras – although conceptually important, are secondary.

This is a good place to recall that, contrary to the above, philosophical structuralists (e. g., Landry [18]) defend the claim that mathematical models are *structures* and category theory is the correct framework for their study. Structuralists have two major arguments against the equational (syntactic) view of mathematical models.

The first reason why the popular syntactic conception of models is troubling to philosophers of science is the following. Even though a stipulated system of equations of a target system admits infinitely many alternative formulations – linked by scale, coordinate, change-of-variable and

other transformations, researchers do not presume of having different models presented by these inessentially different formulations. It is important to bear in mind that the structure of the associated *solution space*, serving as a storehouse for all model-based information about the target system’s causal mechanisms and behavior, remains basically the same. Each type of *solution* of a given system of differential (difference) equations is not defined by the system of equations *per se*, but generally by a much larger (prime) *differential* (difference) *ideal* of equations, generated by it. Since solutions and their mathematical semantics are more important than change-of-variable transformations, systems of equations are related to each other in a considerably deeper way by solution-preserving transformations than by, say, scale transformations. A case in point is the familiar physical equivalence of Hamiltonian and Lagrangean formulations of equations of motion in mechanics. Though the underlying language in each case is different, nevertheless the modeling results are the same. In brief, models in the sense suggested by the practice of science possess properties equations do not seem to have.

It is well known that the basic linear second-order pendulum differential equation  $d^2\theta/dt^2 + g/\ell\theta = 0$  can be put in the form of two linear *first-order* differential equations

$$v = \frac{d\theta}{dt} \quad \text{and} \quad \frac{dv}{dt} + \frac{g}{\ell}\theta = 0,$$

illustrating the notion of an equation-based *state space model*. As an aside, we note that this type of transformation is general in that any system of higher-order differential equations can be rewritten as a first-order system of higher dimensionality.

In this system of equations, the position-velocity pairs  $\langle\theta(t), v(t)\rangle$  are thought to encode the target pendulum’s *dynamical state* (or physical mode of being) of interest at time  $t$ . More importantly, the equations’ smooth, two-dimensional parametric solutions are now given by the position-velocity function pair

$$\begin{aligned} &\langle\theta_{a,b}(t), v_{a,b}(t)\rangle \\ &= \left\langle a \cos(\omega t) + \frac{b}{\omega} \sin(\omega t), b \cos(\omega t) - a\omega \sin(\omega t) \right\rangle, \end{aligned}$$

with parameters  $a = \theta(0)$  and  $b = v(0)$ . Thus, to specify a particular (position-velocity) state-space trajectory for the pendulum model, all we need is the *initial state*  $\langle a, b \rangle$ , arrived at via first and second integration of the original equation. We mention in passing that if the differential equation characterizing a target system were of order  $n$ ,

then generally we would need  $n$  independent parameter values for a unique individuation of any of its solutions in the form of specific time-functions.

The point of this simple exercise in pendulum modeling is to make a headway in the direction of a powerful alternate *structuralist* formulation of the pendulum model that has the following *smooth group action* (smooth flow or smooth group representation) form

$$\mathbb{T} \times (\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R}) \xrightarrow{\langle \theta, v \rangle} \mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R},$$

to be defined next. Here and below,  $\mathbb{T}$  denotes a continuum-time domain, isomorphic to the group  $\langle \mathbb{R}, 0, + \rangle$  of real numbers, and  $\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R}$  denotes the two-dimensional cylinder-shaped state space, generated by the unit circle and intended to encode all physically relevant states in terms of values of the parametrized position-velocity function. The customary geometric interpretation of state-parametrized solutions consists of smooth curves, covering the cylinder state space and thereby forming the target system's *phase portrait*.

By a *continuous group action* of a topological group  $\langle \mathbb{T}, 0, + \rangle$  (interpreted as a continuum-time domain or time-group) on a topological space  $X$  (thought of as a *state space*) we mean a jointly continuous map of the form  $\delta: \mathbb{T} \times X \rightarrow X$  (also known as a dynamical *transition map*) such that the following axioms of *group action* hold for all time instants  $t, t' \in \mathbb{T}$  and for all states  $x$  in  $X$ :

- (i) *Identity property*:  $\delta(0, x) = x$ , and
- (ii) *Group property*:  $\delta(t, \delta(t', x)) = \delta(t + t', x)$ .

Because in what follows, group actions will be used mainly to represent the temporal evolution of target systems' states, the above-introduced group action  $\delta: \mathbb{T} \times X \rightarrow X$  is alternatively called a (deterministic) *topological dynamical model* and is symbolized more succinctly by the curved group-action arrow  $\mathbb{T} \curvearrowright_{\delta} X$ . An impressively large class of deterministic dynamical models arises from autonomous systems of ordinary first-order differential equations by simple state-parametrizations of their solution spaces. If the time domain  $\mathbb{T}$  and the state space  $X$  are both smooth manifolds, and if the transition map  $\delta$  is also smooth (i. e., infinitely differentiable), then  $\mathbb{T} \curvearrowright_{\delta} X$  is called a (deterministic) *smooth dynamical model*. In particular, the structural variant of the earlier discussed smooth pendulum dynamical model has the form  $\mathbb{T} \curvearrowright_{\delta, v} (\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})$ , obtained directly from the solutions of a pair of first-order pendulum equations. Because the time group's operation is an action on its own underlying space, we automatically obtain the 'clock' dynamical model  $\mathbb{T} \curvearrowright_{+} |\mathbb{T}|$ .

Passage from equations to group action formulations of their solutions represents a significant change of viewpoint. From this new *structuralist* perspective, in formulating a state-based mathematical model or simply a *dynamical model* of a target system's behavior, the modeler needs to specify the following two conceptual ingredients: a state space  $X$  (endowed with additional smooth, topological, or measurable structure, assumed to be respected by all maps on it) and a time-group action  $\delta: \mathbb{T} \times X \rightarrow X$ , satisfying the identity and group properties. A defining property of a state space is its being the domain of real-valued, observable quantities. A prime example of an observable quantity in the pendulum model is the Hamiltonian, representing the pendulum's energy levels.

Structuralists argue that the particular system of equations (if any) that generates the group action, is of secondary concern. In a dynamical model, comprised of a state space and a time-group action on it, each state-space point is an abstract, information-bearing encoding of the target system's physical mode of being at a given time. Assuming that the system under consideration has a physical state structure, its adequate state space model comes with a state space comprised of points that are presumed to contain *complete information* about the past history of the system, relevant to its future behavior. Thus, if the target system's instantaneous physical state – encoded by a point in its representing state model, is known, then the system's subsequent temporal evolution can be predicted with the help of the model's time-group action, without any additional knowledge of what has happened to the system. The group action need not be continuous or smooth. It can also be measurable, computable, discrete, and local (i. e., only partially defined).

The structuralist point of view of mathematical models also includes the so-called *behavioral* approach, proposed by Jan Willems [35]. It turns out to be just as effective as the group action approach. Recall from the basic pendulum model example above that the group action pair  $\langle \theta, v \rangle$  has a function space transpose

$$\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R} \xrightarrow{\widehat{\langle \theta, v \rangle}} (\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})^{\mathbb{T}},$$

defined by the time function  $[\widehat{\langle \theta, v \rangle}(\langle a, b \rangle)](t) =_{\text{df}} \langle \theta_{a,b}(t), v_{a,b}(t) \rangle$  for all time instants  $t$  and initial states  $\langle a, b \rangle$  in  $\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R}$ . Note that here the image function space  $\widehat{\langle \theta, v \rangle}(\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})$  is comprised of those time functions (pictured as state trajectories) in  $(\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})^{\mathbb{T}}$  that satisfy the initially given pair of first-order pendulum equations.

In engineering applications, it is often conceptually advantageous to work directly with the solution space

$$\left\{ \theta \mid \frac{d^2\theta}{dt^2} + \frac{g}{\ell}\theta = 0 \right\} \subseteq C^\infty(\mathbb{T})$$

of higher-order (linear) equations – interpreted as the subspace of *behaviors*, without any passage to a parametrizing state space. In this behavioral setting, a *behavioral topological dynamical model* is a triple  $\langle \mathbb{T}, X, \mathcal{B} \rangle$ , consisting of a continuum-time domain  $\mathbb{T}$ , a topological space  $X$  (interpreted as a *signal* space), and a function subspace  $\mathcal{B} \subseteq X^{\mathbb{T}}$  of maps (encoding behaviors), satisfying the system's dynamical equations. (Here and below,  $X^{\mathbb{T}}$  denotes the space of all continuous time-functions, i.e., continuous maps from  $\mathbb{T}$  to  $X$ , endowed with the product topology.) This notion of a model is an outgrowth of the traditional input-output method, popular in engineering. The proponents of the behavioral approach in algebraic systems theory emphasize that in general the space of time-functions  $X^{\mathbb{T}}$  need not be just the space  $C^\infty(\mathbb{T})$  of all smooth functions. Instead, it can also be the space of compactly supported smooth functions, locally integrable functions, distributions, and so forth. Polderman and Willems [27] provide a very readable introduction to the behavioral conception of mathematical models in systems theory. The foregoing lengthy digression into structuralist conceptions of models completes the first reason why the syntactic definition of mathematical models is unsatisfactory. We are now ready for the second reason.

The second reason why the equational conception in science is inadequate is because equations are unable to deal fully and directly with intended empirical interpretations, representational power, denotation, model networks, and many other semantic and representational functions of models. To place the equations on a mathematically rigorous ground, the model builder must choose a *host* ring or field for the possible values of indeterminates and observable quantities. Beyond that, several other choices have to be made, including the ‘correct’ choice of a function space for indeterminates and quantities, and that of parameter spaces for constants. Depending on the interpretation, solutions can be everywhere defined and smooth, or they can be generalized functions (distributions), and so forth. There is also the problem of parameter identification and the question of how various quantities are to be measured. These functions of models are typically determined by the intended interpretation of solution spaces and phase portraits. Last but not least, parent models come with various structural enrichments and impoverishments – forming a model network. But of course

in many applications, mathematical models involve both linguistic and structural (non-linguistic) elements. In representing natural systems, usually equations provide an all-important starting point. However, in view of various inevitable abstractions and idealizations, the corresponding group-action and behavioral models – having “more meat”, serve uniquely well as equation-world *mediators* or *proxies*, and *stand-ins* for the real-world target systems. Indeed, statements derived from well-confirmed equations are strictly true in the associated group-action and behavioral models, but they are only indirectly and approximately true of the actual system's behavior. The applied mathematics literature usually leaves implicit many of the details needed for understanding the equations.

So what is the merit of philosophical objections to the equational treatment of mathematical models? In their seminal work, Oberst [26], Röhrli [28], and Walcher [34] have established a natural *Galois correspondence* between the two (syntactic and structural) approaches, leading to a deeper understanding of the relations between the properties of (differential and difference) equations and the properties of their solutions, representing behavior. This fundamental relationship between the world of equations and that of solution spaces (or input-output behaviors) has been established for a large class of linear (and also for some nonlinear) cases, in the form of category-theoretic duality. In parallel with the above, classical models of statistical experiments, decision theory, game theory, and so forth, built over underlying smooth, topological, metric and measurable spaces, possess computationally convenient algebraic counterparts, granted by Stone–Gelfand duality results, discussed, e.g., in [17].

Having studied the ways in which equations lead to group-actions on state spaces and function spaces of trajectories representing behaviors, we now discuss how these and other types of models can be put together from more basic mathematical structures, and how certain universal constructions support the construction of complex models from simple ones.

### Philosophical and Mathematical Structuralism

Since much of what is taken as distinctive of mathematical models and their connections to target systems is tied to mathematical structuralism, a brief description of the concept of structure is in order. We begin by reviewing the principal theses of *philosophical structuralism*. Within the current philosophical literature (e.g., [8]) on the status of scientific realism, two major positions can be discerned. The first, resurrected by Worrall [36], is *epistemic structuralism* that places a special restriction on scientific



knowledge. Its central thesis is that we can have knowledge of structures without knowledge of natures. Here structures are identified with relations – broadly understood, and natures are the intrinsic modes of being of objects that are related. This view is motivated by the need to handle the ontological discontinuity across theory-change in terms of a well-grounded knowledge of structures. For example, although conceptions of the nature of light have changed drastically through history (from a particle ontology to a wave ontology, and then again to a wave-particle duality), nevertheless, the old equation-based models describing the propagation of light have not been abandoned. Rather, at each stage of knowledge, they have been incorporated into more refined successor models. The second position, called *ontic structuralism*, initiated by Ladyman [16], holds that since structure is all there is to reality, what we can have at best is knowledge of the structural aspects of real-world systems. Inspired by the peculiar nature of quantum theory, ontic structuralism argues for the thesis that there are no objects. This object-free ontology leaves its proponents with the burden of showing how physico-chemical entities (e.g., superconductors and radioactive chemicals), endowed with astonishing causal powers, can be dissolved into structures, without anything left over.

In the field of mathematical ontology, structuralism is regarded as one of the most successful approaches to the philosophy of mathematics. For example, Shapiro [31] states that mathematics is the study of independently existing structures, and not of collections of mathematical objects per se. This picture may seem to be a bit distorted to some, since not all of mathematics is concerned with structures. For example, the distribution of prime numbers in the set  $\mathbb{N}$  of natural numbers and arguments as to why  $\pi$  must be a transcendental real are hardly structural matters.

Be that as it may, from the standpoint of classical model theory, a mathematical structure is simply a list of operations and relations on a set, together with their required properties, commonly stated in terms of equational axioms. The most familiar examples of such structures are Bourbaki's *mother structures* on sets: *algebraic* (e.g., time groups and quantity rings), *topological* (e.g., metric, measurable and topological state spaces), and *order-theoretic* (e.g., partially ordered sets of observables and event algebras). Of course, there are many vastly more elaborate *composite* types of set-theoretic structures, including ordered groups, topological groups, ordered topological groups, group actions, linear spaces, differentiable manifolds, bundles, sheaves, stacks, schemes, and so forth. The world of sets is granted by a fixed maximal (possibly Pla-

tonist) mathematical ontology out of which all structures of mainstream mathematics are presumed to be built.

The foregoing idea of mathematical structure is due mainly to the influence of Nicolas Bourbaki [7], who (together with other mathematicians around him) noted that mathematical structures of practical interest can be generated by three universal operations on sets: The (Cartesian) *product*  $X \times Y$  of two sets  $X$  and  $Y$ , the *power set*  $\mathfrak{P}(X)$  of  $X$ , and the exponentiation-type *function set*  $Y^X$  (consisting of all functions that map  $X$  into  $Y$ ). For example, a group structure on a set  $X$  can be viewed as a designated element of the function set  $X^{(X \times X)}$  with the usual equational properties of the group operation, together with a suitable element of  $X^X$  (for the inverse operation) and a designated element  $0 \in X$ . Similarly, a topological structure on a set  $X$  is given by a designated element of the iterated power set construct  $\mathfrak{P}(\mathfrak{P}(X))$ , satisfying the axioms of topology. Thus, Bourbaki's concept of a mathematical structure is given by an underlying set together with some higher-order set-theoretic data, forming a string of designated elements of suitable product, power or function set constructions on it, and satisfying certain axioms. Unfortunately, Bourbaki's definition of the concept of mathematical structure includes many pathological examples of no known utility. Furthermore, in general, Bourbaki does not consider the notion of structure-preserving mappings between mathematical structures, except isomorphisms that are always suitable functions, treated as subsets of the Cartesian product of their domains and codomains.

The plurality of set theories with divergent properties in particular and that of universes of mathematical objects in general with wildly different and mutually inconsistent formulations have led to new ideas in philosophical structuralism. Among other things, these ideas suggest that in the presence of an abstract product  $\times$ , exponentiation  $(\cdot)^{(\cdot)}$  and other universal operations, any entity whatever can serve as a proxy for the underlying set-theoretic domain or carrier of a structure. This undermines the Platonist commitment to sets, believed to underly mathematical reality. The plurality of set theories has been replaced by a plurality of *toposes* – a special class of categories with products, function space constructions and subobject classification, in which most set-like mathematical constructions can be carried out. Category theory is seen by many as providing a structuralist framework for mathematics and hence also for mathematical models per se. Simply, structures are determined up to isomorphism, making the particular nature, individuality or constitution of their underlying domains irrelevant. Domains are only 'positions' in structured systems. For example, real numbers are fully specified by the structure of a complete Archimedean ordered

field that has many realizations. Mathematics is not about objects, either empirical or mathematical. It is about the axiomatic presentation of the structure of such objects in general, and of no objects in particular. If mathematics discusses objects, it does so only by construing them as ‘positions’ in structured systems, devoid of any special ontology. In sum, mathematical objects need not be sets and mappings need not be functions. However, mathematical structuralism does not always translate into physical structuralism. For example, the quantum structure of particles and the physical structure of liquids is empirically meaningless when viewed in isolation from their concrete physical carriers.

Present-day mathematics uses structure-preserving mappings to specify different kinds of sets-with-structure. For example, instead of describing groups traditionally in terms of elements and higher-order data (i.e., operations satisfying the group axioms) on a set, mathematical practice demonstrates that it is far more effective to treat groups as abstract black-box type objects together with distinguished maps, emulating group homomorphisms between them. Upon engaging product and other operations in the construction of these abstractly given objects and maps, algebraists quickly arrive at a crucial universe of discourse, namely the category **Grp** of groups and group homomorphisms. For a thorough discussion of categories, see [22].

In [30] Dana Schlomiuk specifies the classical notion of a topological structure strictly in terms of abstract spaces and maps of a category **Top** – called the category of topological spaces, independently of the usual set-based means of specification, employing elements and closure operators, or algebras of open or closed subsets. Even far more sophisticated structures can be treated in this fundamentally structuralist manner. For example, a smooth structure of a differentiable manifold can be given category-theoretically by specifying exactly which continuous maps must be smooth in the category of abstractly conceived smooth manifolds. Here the utility of category theory is in providing a uniform treatment of the concept of structure, solely in terms of holistically given ‘structure-preserving’ maps that serve remarkably well also the needs of empirical applications.

Having now indicated the ways category theory specifies mathematical structures, we may profitably revisit the earlier discussed syntactic *equational* and *structuralist* approaches to mathematical models from the point of view we have just developed. Here we only give one illustration of this approach out of many possibilities, namely the category **Top**<sup>T</sup> of topological dynamical models in the form of a topological group-action on state

spaces by a designated time group  $\langle \mathbb{T}, 0, + \rangle$ , discussed in the previous Section. This category consists of topological dynamical models of the form  $\mathbb{T} \curvearrowright_{\delta} X$ , serving as its objects, where the map  $\delta: \mathbb{T} \times X \rightarrow X$  is a continuous action of time-group  $\mathbb{T}$  on its codomain topological state space  $X$ . Next, maps between dynamical models of the form  $\varphi: (\mathbb{T} \curvearrowright_{\delta} X) \rightarrow (\mathbb{T} \curvearrowright_{\delta'} X')$ , called *dynamorphisms*, are given by continuous mappings  $\varphi: X \rightarrow X'$  between the underlying state spaces such that the diagram

$$\begin{array}{ccc} \mathbb{T} \times X & \xrightarrow{\delta} & X \\ 1_{\mathbb{T}} \times \varphi \downarrow & & \downarrow \varphi \\ \mathbb{T} \times X' & \xrightarrow{\delta'} & X' \end{array}$$

commutes, i.e., the compositions of constituent maps along both paths in the diagram give the same map. In other words, we have the equality  $\varphi \circ \delta = \delta' \circ (1_{\mathbb{T}} \times \varphi)$ . Because the identity maps of state spaces are trivially dynamorphisms and since the composition of two dynamorphisms is again a dynamorphism, by definition, **Top**<sup>T</sup> is indeed a category.

A dynamorphism  $\varphi$  from  $\mathbb{T} \curvearrowright_{\delta} X$  to  $\mathbb{T} \curvearrowright_{\delta'} X'$  is said to be an *isomorphic dynamorphism* or simply a *conjugacy* provided that  $\varphi: X \rightarrow X'$  is a homeomorphism. From the standpoint of mathematical systems theory, two conjugate dynamical models are structurally identical. Any dynamical property (i.e., a property defined in terms of a group action on its underlying state space) possessed by one is also possessed by the other. Concretely, a dynamical model  $\mathbb{T} \curvearrowright_{\delta} X$  is said to be *structurally stable* provided that any dynamical model  $\mathbb{T} \curvearrowright_{\delta'} X$  with transition map  $\delta'$  sufficiently close to  $\delta$  in a suitable topological sense is conjugated to  $\mathbb{T} \curvearrowright_{\delta} X$ .

Surjective (onto) dynamorphisms are called *factor* dynamorphisms. Each factor dynamorphism  $\varphi: X \rightarrow X'$  comes with its natural equivalence relation, given by  $x_1 \equiv x_2$  iff  $\varphi(x_1) = \varphi(x_2)$ , and induced *quotient* dynamical model  $\mathbb{T} \curvearrowright_{\delta_{\varphi}} X/\equiv$ . Injective (one-to-one) dynamorphisms specify dynamical *submodels*. For example, the trajectory passing through a state  $x$  at time zero, defined by  $\mathbb{T}(x) = \{\delta(t, x) \mid t \in \mathbb{T}\}$ , can be viewed as a member of the family of the smallest nontrivial submodels of  $\mathbb{T} \curvearrowright_{\delta} X$ . Many other notions and constructions available in the category **Top** of topological spaces automatically transfer to the category **Top**<sup>T</sup> of topological dynamical models and dynamorphisms. For example, the *product*  $(\mathbb{T} \curvearrowright_{\delta} X) \times (\mathbb{T} \curvearrowright_{\delta'} X')$  of two dynamical models  $\mathbb{T} \curvearrowright_{\delta} X$  and  $\mathbb{T} \curvearrowright_{\delta'} X'$  is defined in the same way as in topology, namely by

$$(\mathbb{T} \curvearrowright_{\delta} X) \times (\mathbb{T} \curvearrowright_{\delta'} X') =_{\text{df}} \mathbb{T} \curvearrowright_{\delta, \delta'} (X \times X'),$$

where the diagonal action is given by  $[\delta; \delta'](t, (x, x')) = \langle \delta(t, x), \delta'(t, x') \rangle$ . Disjoint sum (coproduct) of two dynamical models is defined similarly. For more details regarding category-theoretic constructions, see [20] that includes an elaborate category-theoretic treatment of dynamical models. Similar constructions work with varying degrees of success in many other host categories. For example, if **Top** is replaced by the category **Mes** of measurable spaces and measurable maps between them, we obtain the much studied category **Mes**<sup>T</sup> of measurable dynamical models. Another significant mutation of dynamical models arises, when the continuum-time domain is replaced by a discrete-time domain, isomorphic to the group  $\langle \mathbb{Z}, 0, + \rangle$  of integers or the monoid  $\langle \mathbb{N}, 0, + \rangle$  of natural numbers.

Topological (smooth or measurable) dynamical models provide complete support also for the representation of *perturbed, controlled, nonautonomous, and random* dynamical systems, in terms of various skew-product constructions. In more detail, let  $\mathbb{T} \curvearrowright_{\delta} X$  be a (topological, smooth or measurable) *base* dynamical model that represents a *driving* system (i. e., perturbation, control or environmental noise process), affecting the target system's dynamical behavior. The model of this behavior, subject to the influence of added perturbation, control, and so on, is based on a so-called (continuous, differentiable or measurable) *cocycle* map  $\alpha: \mathbb{T} \times X \times Y \rightarrow Y$ , acting on the *driven* target system's principal state space  $Y$  and satisfying the following *skew-product action* (or *cocycle*) axioms for all  $t, t'$  in  $\mathbb{T}$ ,  $x \in X$ , and  $y \in Y$ :

- (i) *Identity property*:  $\alpha(0, x, y) = y$ , and
- (ii) *Cocycle property*:  $\alpha(t + t', x, y) = \alpha(t, \delta(t', x), \alpha(t', x, y))$ .

The space  $X \times Y$  should be thought of, informally, as a *trivial bundle*, made up of fibers  $\{x\} \times Y$ , indexed by points  $x \in X$  and “glued together” by the topology. In the same way, for time instants  $t, t'$ , the cocycle map should be viewed as an indexed family

$$\{x\} \times Y \xrightarrow{\alpha(t, \bullet, \cdot)} \{\delta(t, x)\} \times Y \xrightarrow{\alpha(t', \delta(t, \bullet), \cdot)} \{\delta(t' + t, x, \cdot)\} \times Y$$

of maps between fiber state spaces. As the extant base state  $x$  at time  $t$  is shifted by the base model's dynamics to  $\delta(t, x)$ , the restricted cocycle map  $\alpha(t, x, \cdot)$  moves each system state  $y$  in the fiber  $\{x\} \times Y$  over  $x$  to the state  $\alpha(t, x, y)$  belonging to the fiber  $\{\delta(t, x)\} \times Y$  over  $\delta(t, x)$ .

The *skew-product*  $(\mathbb{T} \curvearrowright_{\delta} X)_{\alpha}^{\times} Y$  of a base dynamical model  $\mathbb{T} \curvearrowright_{\delta} X$  and a principal state space  $Y$  under cocycle  $\alpha$  acting on  $Y$  is defined by the dynamical model

$$(\mathbb{T} \curvearrowright_{\delta} X)_{\alpha}^{\times} Y =_{\text{df}} \mathbb{T} \curvearrowright_{\phi} (X \times Y)$$

on the trivial bundle  $X \times Y$ , where  $\phi: \mathbb{T} \times (X \times Y) \rightarrow X \times Y$  is specified by  $\phi(t, (x, y)) =_{\text{df}} \langle \delta(t, x), \alpha(t, x, y) \rangle$  for all time instants  $t$  and states  $x, y$ . It is easy to check that the transition map  $\phi$  is fiber-preserving and hence a bundle morphism. Since each group-action induces a skew-product action, product dynamical models are special cases of skew-product dynamical models. More importantly, because the second projection map  $\pi_Y: (\mathbb{T} \curvearrowright_{\delta} X)_{\alpha}^{\times} Y \rightarrow (\mathbb{T} \curvearrowright_{\delta} X)$  is a factor dynamomorphism, skew-products are best viewed as objects of the so-called slice category **Top** <sub>$\mathbb{T} \curvearrowright_{\delta} X$</sub>  that fuses topological (smooth, measurable, etc.) bundle theory with the theory of dynamical systems. *Rohlin's representation theorem* states that the domains of factor morphisms in the category of probability spaces and maps are essentially skew-products. Thus, all stochastic representations of perturbed systems are tractable uniformly by skew-product constructions.

*Nonautonomous* differential equations (describing nonautonomous systems) of the form  $dy/dt = f(t, y)$  with  $t \in \mathbb{T}$  and  $y \in Y$  (involving an explicit time dependence), are known to have solutions that induce skew-product dynamical models of the form  $(\mathbb{T} \curvearrowright_{+} |\mathbb{T}|)_{\alpha}^{\times} Y$  over a suitable cocycle  $\alpha$ . Quite simply, the model enlarges the principal state space  $Y$  by the space  $|\mathbb{T}|$  of time coordinates. Because solutions of stochastic differential equations have the form of cocycles over base measurable dynamical models, once again, smooth skew-products, called *random dynamical models*, provide the correct structuralist framework for them. For a thorough treatment of some of these topics, see [1] and [9].

### Three Approaches to Applying Mathematical Models

Since one of the most perplexing philosophical questions is how models relate to their target systems, in this Section we review three major approaches to the problem of model-world relationship.

#### Internal Approach

Set-theoretic models are applicable to real-world systems simply because physical objects can literally be members of legitimate sets and there are special functions between sets comprised of physical objects (e. g., particles and bodies) and pure mathematical objects (e. g., reals). As a classical example of a physical application of this nature, recall the application of Newton's laws to the solar system. The proponents of the internal view (including, e. g., Steiner [32]) argue that planets together with the sun form a perfectly meaningful finite set on which real-valued mass and position functions can be defined in the usual way. In this man-

ner, set-theoretic models are applied to physics (and other disciplines) using only *internal* (i. e., internal to set theory) relations between, say, bodies and numbers. The defenders of the internal view treat the names of (say) planets as *rigid designators*, always picking the same object in every possible situation in which they exist at all. It is now easy to see how the intended empirical interpretation of mathematical models leads directly to claims about the physical world.

It must be stressed, however, that there is a good basis for questioning this fairly popular approach. One obvious concern is this: For particles, bodies and other physical objects to form classical sets, it must be assumed that they can never be annihilated, split or changed in some other ways, leading to a loss of their identity. Under these types of implicit comprehensive idealizations, it is better to think of such ‘physical’ or ‘empirical sets’ as ordinary sets involving mathematical encodings. Simply, the actual physical objects under study are encodable in terms of suitable abstract elements (e. g., numbers) of a set, so that the informal semantic assumptions of the sort “ $\mathcal{B}$  is a set of physical bodies”, “ $X$  is a set of physical states”, “ $\mathbb{T}_\delta X$  is a dynamical system”, and so forth, are treated as succinct abbreviations of the formal details of fallible mathematical encodings and their empirical interpretation. A more serious philosophical concern pertains to the nature of (e. g., differentiable) maps on various ‘empirical sets’, consisting of fluids and other objects of continuum mechanics.

### External Approach

Some philosophers suggest to draw a thick conceptual line between timeless mathematical models and actual physical systems. But then, for a model to be a model of something else, a relationship between a model and what it models is required. Specifically, applications of mathematical models involve certain external (preferably intensional) relations between models and real-world systems. In more detail, philosophical structuralists argue that the model-world relation is best described by structure-preserving maps (usually isomorphisms or embeddings) between representing mathematical models and their intended physical (or generally empirical) situations. As a concrete illustration of this viewpoint, consider the representational model of weight measurement, studied in great detail by Kranz et al. in [15]. The empirical domain  $\mathcal{D}$  of the ‘physical structure’ of weight measurement is a set of material objects. This domain is equipped with a binary ‘heavier-than’ weak-order relation  $\succ$ , where the empirical statement  $d \succ d'$  – stating that object  $d$  is heavier than object  $d'$ , is operationally established by placing  $d$  and  $d'$

on the two pans of an equal-arm balance scale and then observing which pan drops. In addition, there is a binary weakly commutative and weakly associative composition operation  $\oplus$  on  $\mathcal{D}$ , where for objects  $d$  and  $d'$  the value  $d \oplus d'$  denotes the composite object obtained by placing both objects in the same pan with  $d$  beneath  $d'$ . Now, here the world-model relation is given by an algebraic homomorphism from the so-called physical *extensive measurement structure*  $\langle \mathcal{D}, \succ, \oplus \rangle$  of weight measurement to the ordered additive structure  $\langle \mathbb{R}, >, + \rangle$  of reals. Such homomorphisms are guaranteed to exist provided that the extensive measurement structure satisfies certain relatively simple axioms, similar to those used in the definitions of Archimedean-ordered semigroups.

We see that for a structure-preserving world-to-model map to exist, numerous idealizing restrictions must be placed on the target physical domain’s structure. For example, in the case of weight measurement, the Archimedean property forces the empirical domain  $\mathcal{D}$  to be infinite, even though in applications experimenters always work with finitely many objects. Because idealizations typically involve known-to-be-false descriptions, ignorance of causes, and deliberate stipulations of objects that do not exist in the actual physical situation, a structure-preserving map can be guaranteed to exist only between the idealization itself and a representing mathematical model, but not necessarily between the actual real-world situation (enjoying unlimited degrees of freedom) and the model.

Finally, one must question the ontological status of the structure-preserving maps (isomorphisms). If these maps go from a mathematical domain to a real-world physical domain or conversely, then their graphs, being subsets of a ‘hybrid set’ of physical-mathematical object pairs, take us back straight to the internal viewpoint, questioned earlier.

A rival and equally popular approach to this general topic is to replace the isomorphism (or homomorphism) relation between models and the world by some other, technically less demanding devices. For example, Giere [13] proposes to solve the problem of structure-preserving maps by passing to considerably simpler graded context-dependent *similarity* relations between models and aspects of the real world. Thus, if the representing model is sufficiently similar to its target system, then the analysis of the model is also (indirectly) an analysis of the system. This solution comes at quite a price, since it is hard to see what exactly, if anything, is similar between the familiar Lotka–Volterra differential equations and the predator-prey two-species system of finitely many sharks and fishes in the Adriatic Sea. Another major concern is that under similarity anything can be a model of anything



else, since any two things can always be claimed to be similar in some respect and to some degree or other. In general, model builders cannot reduce the problem of application of models to similarity relations, because models usually come with falsifying idealizations that simply violate any similarity relation that may perhaps exist between intractable ‘perfect’ models and their systems. Last but not least, mathematical models tend to possess a ‘surplus structure’ and features that do not correspond to anything in particular, similar or not, in the actual system (e.g., recall the applications of complex analysis to the behavior of electric circuits or the role of undetectable measure-zero attractors in a dynamical model).

### Model Network Approach

The relation between a mathematical model and its target system is not as straightforward as it is often thought to be. Models interact with other models, both globally and locally in complex ways, and their relationship to the world is not a static affair. If connections between models and their target systems were tractable by rigid relations, then a large bulk of scientific research would become redundant. However, as mathematical models are structurally mutated, their representational capacities often undergo unexpected refinements and changes. As a result, in subsequent applications new epistemic links between models and systems emerge, and accrue along the way. Thus, from the standpoint of scientific practice, an essential aspect of mathematical modeling is extendability, enrichment and open-endedness of model constructions. Model builders cannot realistically constrain models to be just single frozen structures. Instead, applied mathematical models are best viewed as complex *networks* of structures with distributed empirical interpretations that guide model users in the network-world interface to gather more informative data and make better predictions. (For related views but motivated by different considerations, see [4] and [5].)

Applications of category theory to mathematical models has opened up the possibility of building new models from old ones via functorial constructions. In Sect. “[Philosophical and Mathematical Structuralism](#)”, we have given a brief introduction to some concepts, intended for the class of topological dynamical models. But there is more. For example, the study of deterministic dynamical systems operating in a chaotic regime – hard to understand in deterministic terms, calls for a passage to *probabilistic* dynamical models. These are obtained by introducing *probability monads* – special functors discussed by Giry [14], on suitable categories of topological and

measurable spaces. In brief, the parent deterministic dynamical model  $\mathbb{T}_{\delta}^{\curvearrowright} X$  is used in conjunction with the associated logically higher-level probabilistic dynamical model  $\mathbb{T}_{\delta^*}^{\curvearrowright} \mathcal{D}(X)$  with time-group action on the convex space  $\mathcal{D}(X)$  of probability density functions, representing the target system’s stochastic states. Here the transition map  $\delta^*$  is defined by the so-called *Perron–Frobenius integral equation*, studied in great detail by Lasota and Mackey [19]. Mutation of  $\mathbb{T}_{\delta}^{\curvearrowright} X$  into nondeterministic, fuzzy and other dynamical models is also obtained by the monad approach. Since the time evolution of states of a target system is observed only indirectly via measurable real-valued quantities, defined on the representing model’s state space  $X$ , the *parent* model  $\mathbb{T}_{\delta}^{\curvearrowright} X$  is used in parallel with another associated higher-level dynamical model  $\mathbb{T}_{\delta^*}^{\curvearrowright} \mathbf{C}(X)$  on the Banach algebra  $\mathbf{C}(X)$  of (bounded) real-valued continuous functions on  $X$ . The induced time-group action  $\delta^*$ , defined by the fundamental *Gelfand duality* result, characterizes the time evolution of observable quantities, crucial in time-series analysis.

Because in applications the evaluation of the analytic parent model’s transition map tends to involve roundoff and other truncation operations that may have long-term pathological effects, and since generally the model’s time and space points cannot be identified with arbitrarily fine degrees of accuracy, several forms of (temporal, spatial and parametric) *discretizations* become necessary. *Time discretizations* of a continuum-time model  $\mathbb{T}_{\delta}^{\curvearrowright} X$  is solved by passing to discrete-time dynamical models  $\tau \mathbb{Z}_{\delta\tau}^{\curvearrowright} X$ , where  $\tau \mathbb{Z} = \{\dots, -2\tau, -\tau, 0, \tau, 2\tau, \dots\}$  is the group of discrete timesteps with sampling period  $\tau > 0$ . *Spatial discretization* of  $\mathbb{T}_{\delta}^{\curvearrowright} X$  is captured by a nested family of coarse-grained dynamical models of the form  $\mathbb{T}_{\delta_{\epsilon}}^{\curvearrowright} X_{\epsilon}$ , where the underlying state space  $X_{\epsilon}$  is comprised of a uniform grid (lattice) of finitely many points in  $X$  with mesh size  $\epsilon = \frac{1}{n}$ , tractable with finitary resources. Spatially discretized dynamical models provide the formal meeting ground for measurement results and the parent model’s predictions. To understand the gap between *ontologically* and *epistemologically* motivated dynamical models, it is important to know how well a discretized space mimics the properties of the parent model’s state space. Remarkably, Hausdorff topological state spaces are known to be approximable by inverse limits of nested sequences of  $T_0$  finite topological spaces. For more details on this subject, see [6].

### Validating Mathematical Models

Although many models are built for scientific research purposes, applications demand that models be validated iteratively by comparisons of their predictions with mea-

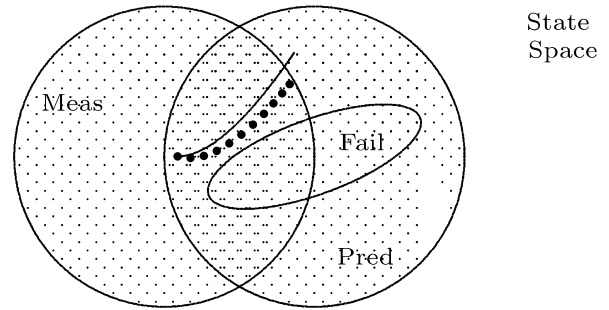


surement data. Validation is intended to demonstrate that the representing model generates predictions and claims in its domain of applicability that are consistent with its intended application – within an acceptable range of accuracy. Because it is not true that good predictions can only be obtained from a model that is causally sound and since data may not portray the target system accurately, full *verification* of a model requires various testing procedures for agreement with cause-effect relationships (operating in the target system) and analysis of inter-model relations in a model network to which it belongs.

Validation tends to be limited by the available domain of measurement data – to be used in model-world comparisons, and generally is not sufficient to demonstrate that the model is adequate in its entire state space, modulo locally granted margins of error. As a matter of fact, mathematical models are seldom valid in all regions of their phase portraits. As we have seen, the most popular example is the Newtonian dynamical model of a simple pendulum that gradually fails as its angular velocity approaches the speed of light. In scientific practice, a dynamical model is typically valid only in limited regions of its underlying state space that are of particular interest. Problems arise when the model fails to be valid in most parts of the state space regions under consideration. Deficiencies in the model may be traced to a wrong choice of differential equations (that ignores higher degrees of nonlinearity), a crude choice of time and/or space discretization parameters, and reliance on data samples contaminated with gross errors.

In many applications, it is far from sufficient to know that the model is *statistically* valid. The reason is that a statistically valid theoretical model may turn out to be *dynamically* (qualitatively) invalid, meaning that it may fail to represent correctly the dynamical invariants of the target system (including equilibrium states, periodic, aperiodic, chaotic or strange attractors, and their basins) under various choices of parameter values. Note that even if a dynamical model provides highly accurate predictions in its tested state space regions, it does not necessarily follow that the model is dynamically valid, since it may include spurious dynamics in untested state space regions that are also of interest. The spectrum of actual behaviors of the target system remains largely unknown from the perspective of its representing model. Further observation and theoretical research may be needed to obtain a better knowledge about the dynamical invariants and characteristics of the system.

These considerations lead us to reason about the validation of dynamical models geometrically, in terms of suitable conditional geometric measures of adequacy.



Philosophy of Science, Mathematical Models in, Figure 1  
Geometry of model validation

Given a parent dynamical model of a target system, suppose the associated body of actual observation and *measurement* data, available at a given stage of knowledge and collected independently of the model under consideration, forms a subset *Meas* of a finite grid of the parent model's underlying state space, as illustrated by a shaded circle in Fig. 1 below. The size of the set *Meas* is constrained by various resources, available measuring instruments, their accuracy, and research interests of the validating scientist. Of course, *Meas* varies with time and it can form a multiply connected or scattered subset of the state space. For simplicity, we assume that the discretization embodied in the grid captures the uniform admissible margin of error.

Along related lines, let *Pred* be a subset of a finite grid of the parent model's state space, given by all quantitative *predictions* calculated from the representing model at a given stage of research and depicted by the second shaded circle in Fig. 1 that overlaps with *Meas*. Naturally, the set *Pred* grows in time as brand-new predictions are generated by the model. It should be obvious that the intersection  $\text{Meas} \cap \text{Pred}$  represents those predictions that have been checked by observation or measurement.

Since the model of interest need not be perfect or fully reliable, it is likely to generate some predictions that falsify the model. In Fig. 1, the set of *falsifiers* is indicated by the ellipse *Fail*, forming a subset of *Pred*. Note that at the current stage of research, only a proper subset of incorrect predictions in *Fail* is comparable with measurement results. The other potentially falsifying predictions in *Fail* have not yet been verified. Finally, note also that subsequent discrete measurements of a state trajectory (indicated in Fig. 1 by a dotted curve segment) over a longer period of time may gradually diverge from the corresponding parent model-given continuous state trajectory, both generated by the same presumed initial state.

We are now ready to use the geometric scheme developed above to *measure the adequacy of dynamical models*.

Even though measurement results and predictions form finite sets, for the sake of simplicity we choose to measure the *adequacy* of the given dynamical model by the ratio

$$\frac{\Lambda(\text{Meas} \cap (\text{Pred} - \text{Fail}))}{\Lambda(\text{Meas})},$$

where  $\Lambda$  denotes the Lebesgue measure defined on the model's underlying state space. The idea is that the ratio of the volume of fully verified *correct* predictions and the volume of all available measurement results is a good indicator of the model's adequacy at a given stage of knowledge. In any case, it should be clear that a *semantic* approach to confirmation theory must directly engage the structure of dynamical models under consideration. Because adequacy and reliability should be judged according to the volume of phase portrait regions on which measurements and predictions agree, modulo admissible errors, measures of the sort displayed above are in the ballpark of measuring adequacy. Note, however, that just like many other measures in dynamical systems theory (including topological and Kolmogorov–Sinai entropy functions), the foregoing measure of model adequacy is largely conceptual and not readily implementable in all instances of real-life models.

Testing models for adequacy is conceptually quite similar to testing statistical hypotheses. Recall that a given representative sample of data validates a statistical hypothesis about the target population only with a certain degree of confidence; the larger the sample, the higher the degree of confidence in the correctness of the hypothesis. Likewise, representative measurement data validate the dynamical model under consideration via its predictions pertaining to the target system's addressed behavior only locally, specified by a region in its phase portrait. The larger the body of validating data in its phase portrait, the higher the degree of confidence in the model's global adequacy.

### Future Directions

We do not give a detailed list but briefly mention two major directions of current research.

### Galois Correspondence

#### Between Equational and Structuralist Approaches

There are several algebraic methods in the realm of linear differential equations that demonstrate a *Galois correspondence* between differential or difference ideals of equations (modules of differential operators) and group-action (behavioral) models. There are reasons to conjecture that these results remain valid also for a large class of *nonlinear* differential or difference equations. Although some ex-

amples of this are studied by Walcher [34], a more complete understanding of the equation-solution relationship is needed. A similar Galois connection must be addressed also in the context of stochastic differential equations and random dynamical models or stochastic flows.

### Validation and Verification of Mathematical Models

It is a problem of considerable interest to formalize the processes of model validation and testing. In recent years efforts have been made to come to grips with this problem in the field of validation of computer simulation models. (See, for example, the influential work of Sargent [29].) However, a substantive theory of dynamical model validation that circumvents the inadequacies of classical confirmation theory is not yet available.

### Bibliography

#### Primary Literature

1. Arnold L (1998) Random dynamical systems. Springer, New York
2. Atmanspacher H, Primas H (2003) Epistemic and ontic realities. In: Castell L and Ischebeck O (eds) Time, Quantum and Information. Springer, New York, pp 301–321
3. Bailer-Jones DM (2002) Scientists' thoughts on scientific models. Perspectives Sci 10:275–301
4. Balzer W, Moulines CU, Sneed J (1987) An Architectonic for science: The structuralist program. Reidel, Dordrecht
5. Balzer W, Moulines CU (1996) Structuralist theory of science. Focal issues, new results. de Gruyter, New York
6. Batitsky V, Domotor Z (2007) When good theories make bad predictions. Synthese 157:79–103
7. Bourbaki N (1950) The architecture of mathematics. Am Math Mon 57:221–232
8. Chakravartty A (2004) Structuralism as a form of scientific realism. Int Stud Philos Sci 18:151–171
9. Colonius F, Kliemann W (1999) The dynamics of control. Birkhäuser, Boston
10. Da Costa NCA, French S (2003) Science and partial truth. A unitary approach to models and scientific reasoning. Oxford University Press, New York
11. Erkal C (2000) The simple pendulum: a relativistic revisit. Eur J Phys 21:377–384
12. Fulford G, Forrester P and Jones A (1997) Modelling with differential and difference equations. Cambridge University Press, New York
13. Giere R (2004) How models are used to represent reality. Philos Sci (Proc) 71:742–752
14. Giry M (1981) A categorical approach to probability theory. In: Banaschewski B (ed) Categorical Aspects of Topology and Analysis, Lecture Notes in Mathematics, vol 915. Springer-Verlag, New York, pp 68–85
15. Krantz DH, Luce RD, Suppes P, Tversky A (1971) Foundations of measurement. Academic Press, New York
16. Ladyman J (1998) What is structural realism? Stud Hist Philos Sci 29:409–424

17. Lambek J, Rattray BA (1979) A general Stone-Gelfand duality. *Trans Am Math Soc* 248:1–35
18. Landry E (1999) Category theory: The language of mathematics. In: Howard DA (ed) *Proceedings of the 1998 Biennial Meeting of the Philosophy of Science Association*. The University of Chicago Press, Chicago, pp S14–S26
19. Lasota A, Mackey MC (1994) *Chaos, Fractals and Noise. Stochastic Aspects of Dynamics*, 2nd edn. Springer-Verlag, New York
20. Lawvere FW (2005) Taking categories seriously. *Theor Appl Categ* 8:1–24
21. Le Cam L (1986) *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York
22. Mac Lane S (1971) *Categories for the working mathematician*. Springer-Verlag, New York
23. Mac Lane S (1996) Structure in Mathematics. *Philosophia Mathematica* 20:1–175
24. Moore C (1991) Generalized shifts: unpredictability and undecidability in dynamical systems. *Nonlinearity* 4:199–230
25. Morgan M (2005) Experiments versus models: New phenomena, inference and surprise. *J Econ Methodol* 12:317–329
26. Oberst U (1990) Multidimensional constant linear systems. *Acta Appl Math* 68:59–122
27. Polderman JW, Willems JC (1998) *Introduction to mathematical systems theory: A behavioral approach*. Texts in Applied Mathematics 26. Springer-Verlag, New York
28. Röhl H (1977) Algebras and differential equations. *Nagoya Math J* 68:59–122
29. Sargent RG (2003) Verification and validation of simulation models. In: Chick S, Sanchez PJ, Ferrin E, and Morrice DJ (eds) *Proceedings of the 2003 Winter Simulation Conference*. IEEE, Piscataway, pp 37–48
30. Schlomiuk DI (1970) An elementary theory of the category of topological spaces. *Trans Am Math Soc* 149:259–278
31. Shapiro S (1997) *Philosophy of Mathematics. Structure and Ontology*. Oxford University Press, Oxford
32. Steiner M (1998) *The applicability of mathematics as a philosophical problem*. Harvard University Press, New York
33. Suarez M (2003) Scientific representation: Against similarity and isomorphism. *Int Stud Philos Sci* 17:225–244
34. Walcher S (1991) *Algebras and differential equations*. Hadronic Press, Palm Harbor
35. Willems JC (1991) Paradigms and puzzles in the theory of dynamical systems. *IEEE Trans Automatic Control* 36:259–294
36. Worrall J (1996) Structural realism: the best of both worlds? In: Papineau D (ed) *The Philosophy of Science*, Oxford University Press, New York, pp 139–165
- Psillos S (1999) *Scientific Realism: How Science Tracks Truth*. Routledge, London
- Suppes P (2002) *Representation and Invariance of Scientific Structures*. CSLI Publications, Stanford University, Stanford

## Physics and Mathematics Applications in Social Science

DIETRICH STAUFFER<sup>1</sup>, SORIN SOLOMON<sup>2</sup>

<sup>1</sup> Institute for Theoretical Physics, Cologne University, Köln, Germany

<sup>2</sup> Racah Institute of Physics, Hebrew University, Jerusalem, Israel

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Some Models and Concepts](#)

[Applications](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Cellular automata** Discrete variables on a discrete lattice change in discrete time steps.

**Ising model** Neighboring variables prefer to be the same but exceptions are possible. The probability for such exceptions is an exponential function of “temperature”.

**Percolation** Each site is randomly either occupied or empty, leading to random clusters. At the percolation threshold for the first time an infinite cluster is formed.

**Universality** Certain properties are the same for a whole set of models or of real objects.

### Definition of the Subject

Herein we introduce the section of this Encyclopedia devoted to Social Sciences, edited by A. Nowak, which concentrates on the application of mathematics and physics to this field. Under “mathematics” we include also all computer simulations if they are not taken from physics; while physics applications include model simulations derived from physics that were applied to social simulations. Thus, obviously there is no sharp border between applications from physics and from mathematics in the sense of our definition. Also social science is not defined precisely. Included are economics and linguistics, but not social insects

### Books and Reviews

- Brin M, Stuck G (2002) *Introduction to Dynamical Systems*. Cambridge University Press, New York
- Hellman G (2001) Three varieties of mathematical structuralism. *Philosophia Mathematica* 9:184–211
- Margani L, Nersessian N, Thagard P (eds) (1999) *Model-based Reasoning in Scientific Discovery*. Kluwer, Dordrecht
- Margani L, Nersessian N (eds) (2002) *Model-based Reasoning: Science, Technology, Values*. Kluwer, Dordrecht
- Morgan M, Morrison M (1999) *Models as Mediators. Perspectives on Natural and Social Science*. Cambridge University Press, Cambridge

or fish swarms, nor human epidemics or demography. It should further be noted that the section of the Encyclopedia on agent-based modeling edited by F. Castiglione also contains articles of social interest.

## Introduction

If mathematical/physical methods are applied to social sciences, a major problem is the mutual lack of literature knowledge. Take for example the Schelling model of racial segregation in cities [1]. Sociologists don't cite the better and simpler Ising model, physicists ignored the Schelling model for decades, and sociologists also ignored better sociology work [2]. For simulations of financial markets, many econophysicists thought that they had introduced Monte Carlo and agent-based simulations to finance, not knowing of earlier work from some forward-looking Nobel laureates in economics [3,4]. For inter-community relations, already 25 centuries ago analogies with liquids were pointed out by Empedokles in Sicily ▶ [Opinion Dynamics and Sociophysics](#) or [5]. More recently, Ettore Majorana [6] around 1940 suggested application of quantum-mechanical uncertainty to socio-economic questions. With emphasis shifted to statistical physics, sociophysics and econophysics became fashionable around the change of the millennium, but continuous lines of research by some physicists had already begun in 1971 [7]. In the same year the Journal of Mathematical Sociology started and published Schelling's model of urban segregation [1], which is a modification of the Ising magnet at zero temperature. The year of 1982 saw the start of two other lines of research by physicists on socio-economic questions [8,9].

Languages have been simulated on computers for decades, while the interest of physicists in this area is more recent [10,11], triggered mostly by a model of language competition [12].

We do not mention chemists since at present they play no major role in this field. However, the 1921 chemistry Nobel laureate F. Soddy [13], to whom we owe the "isotope" concept, had already worked on economic, social and political theories, and his finance work of the 1930s was still being cited in 2007. The present authors are trying this the other way round: First apply physics to social sciences, and then get the Nobel prize (for literature: science fiction).

## Some Models and Concepts

Physicist Albert Einstein said that models should be as simple as possible, but not simpler. In this spirit we now introduce some basic physics models and concepts for readers from social sciences. All models are complex in

the sense that the behavior of large systems cannot be predicted from the properties of the single element.

## Cellular Automata

Mathematicians denote cellular automata often as "interacting particle systems", but since many other models or methods in physics use interacting particles, we do not use this term here. A large  $d$ -dimensional lattice of  $L^d$  sites carries variables  $S_i$  ( $i = 1, 2, \dots, L^d$ ) which can be either zero or one; more generally, they are small integers between 1 and  $Q > 2$ . The lattice may be square (four nearest neighbors), triangular (six nearest neighbors), or simple cubic (also six nearest neighbors, but in  $d = 3$  dimensions); many other choices are also possible. Time  $t = 1, 2, \dots$  increases in steps. At each time step, each  $S_i(t+1)$  is calculated anew, one  $i$  after the other, from a deterministic or probabilistic rule involving the neighboring  $S_k(t)$  of the previous time step. This way of updating is called "simultaneous" or "parallel"; one may also use sequential updating where  $S_i$  depends on the current values of the neighbors  $S_k$ ; then the order of updating is important: random sequential, or regular like a typewriter.

An example is a biological infection process: Each site  $i$  becomes permanently infected,  $S_i = 1$ , if at least one of its nearest neighbors is already infected. (Computers handle that efficiently if each computer word of, say, 32 bits stores 32 sites, and if then 32 possible infections are treated at once by bit-by-bit logical-OR operations [14].)

## Temperature

We know temperature  $T$  from the weather reports, but in physics it enters according to Boltzmann into the probability

$$p \propto \exp(-E/k_B T) \quad (1)$$

to observe some configuration with an energy  $E$ . Here  $T$  is the temperature measured in Kelvin (about 273 + the Celsius or centigrade temperature), and  $k_B$  the Boltzmann constant relating the scales of energy and temperature. For simplicity we now set  $k_B = 1$ , i.e., we measure temperature and energy in the same units. If  $g$  different configurations have the same energy, then  $S = \ln(g)$  is called the entropy, and the probability to observe this energy is  $\propto g \exp(-E/T) = \exp(-F/T)$  with the "free energy"  $F = E - TS$ .

In a social application we may think of peer pressure or herding: If your neighbors drink Pepsi Cola, they influence you to also drink Pepsi, even though at present you drink Coca Cola. Thus, let  $E$  be the number of nearest neighbors drinking Pepsi Cola, minus the number of

Coke-drinking nearest neighbors. The probability for you to switch then is given by the energy difference and equal to  $\exp(-2E/T)$  (or 1 if  $E < 0$ ) in the Metropolis algorithm, or  $1/(1 + \exp(2E/T))$  in the Glauber or Heat Bath algorithm. In both cases there is a tendency to decrease  $E$ . In the limit  $T = 0$  one never makes a change which increases  $E$ , while for small positive  $T$  one increases  $E$  with a low but finite probability. In the opposite limit of infinite temperature, the energy becomes unimportant and all possible configurations become equally probable. Neither zero nor infinite temperature are usually realistic.

In this sense, decreasing the energy  $E$  is the most simple or most plausible choice, and the temperature measures the willingness or ability to deviate from this simplest option, e. g. to withstand peer pressure. But temperature also incorporates all those random accidents of life that influence us but are not part of the social model. For example, it may happen that there is no Pepsi Cola available even though all your neighbors drink Pepsi and you want to follow them. Investors have to make their financial choices under the influence of their clients, whose life is shaped by births, marriages, deaths, or other personal events which are not included explicitly into a financial market model. These accidents are then simulated by a finite temperature, entering the probability that one does not follow the usual rule.

The ability to withstand peer pressure and the randomness of personal lives are in principle two different things, and if one wants to include them both one needs two different temperatures  $T_1$  and  $T_2$  [15], which do not exist in traditional physics.

### Ising Model

In the model published by Ernst Ising in 1925, the variables  $S_i$  are not 0 or 1, but  $\pm 1$ :

$$E = - \sum_{i,k} S_i S_k - B \sum_i S_i \quad (2)$$

and for  $B = 0$  this corresponds to the above Coke versus Pepsi example. The first summation runs over all neighbor pairs, the second over all sites. Thus, if site  $i$  considers changing its variable, the energy change is  $\pm \Delta E = 2(\sum_k S_k - B)$  and enters through  $\exp(-\Delta E/T)$  into the probabilities to flip  $S_i$ ; now  $k$  runs over the neighbors of  $i$  only. (If instead of flipping one  $S_i$  one wants to exchange two different variables  $S_i$  and  $S_j$ , moving  $S_i$  into site  $j$  and  $S_j$  into site  $i$ , then one has to calculate the energy changes for both sites  $i$  and  $j$  in this “Kawasaki” kinetics.) A computer program and pictures from its application are

given elsewhere in this Encyclopedia ► [Opinion Dynamics and Sociophysics](#) or in [5].

In physics, the  $S_i$  are magnetic dipole moments of the atoms, often called spins, and  $B$  is proportional to the magnetic field. Usually, physicists write an exchange constant  $J$  before the first sum, but we set  $J = 1$  for simplicity here. The model was invented to describe ferromagnetism, like in the elements iron, cobalt or nickel. Later it was found to describe liquid–vapour equilibria and other phase transitions. We know that iron at room temperature is magnetic, and this corresponds to the fact that for  $0 < T < T_c$  and under zero field  $B$ , the majority of its spins point in one direction (either mostly +1 or mostly −1), while for  $T > T_c$  half of the spins point in one and the other half in the opposite direction. The magnetization  $M = \sum_i S_i$ , often normalized by the number  $L^d$  of spins, is therefore an order parameter. The critical temperature  $T_c$  is often named after Pierre Curie.

In one dimension, we have  $T_c = 0$ ; in the square lattice in two dimensions we know  $T_c = 2/\ln(1 + \sqrt{2})$  exactly, while on the simple cubic lattice  $T_c \simeq 4.5115$  is estimated only numerically. Of course, one has generalized the model to more than nearest neighbors, to more than two states  $\pm 1$  for each spin, and to disordered lattices and networks.

### Percolation

Simpler than the Ising model but less useful is percolation theory, reviewed more thoroughly in this Encyclopedia in the section edited by M. Sahimi. Each site of a large lattice is randomly occupied with probability  $p$ , empty with probability  $1 - p$ , and clusters are sets of occupied neighboring sites. There is one percolation threshold  $p_c$  such that for  $p < p_c$  only finite clusters exist, for  $p > p_c$  also one infinite cluster, and at  $p = p_c$  several infinite clusters may co-exist, which are fractal: The number of occupied sites belonging to the infinite clusters varies at  $p_c$  as  $L^D$  where  $D$  is the fractal dimension. Here “infinite” means: spanning from one end of the sample of  $L^d$  sites to the opposite end, or: increasing in average number of sites with a positive power of  $L$ . In one dimension, again one has no phase transition ( $p_c = 1$ ), on the square lattice  $p_c \simeq 0.5927462$  and on the simple cubic lattice  $p_c \simeq 0.311608$  are known only numerically, with a fractal dimension of 1, 91/48 and  $\simeq 2.5$  in one to three dimensions.

In the resulting disordered lattices, each site has from 0 to  $z$  neighbors, where  $z$  is the number of neighbors in the ordered lattice  $p = 1$ . If one neglects the possibility of cyclic links one finds  $p_c = 1/(z - 1)$  in this Bethe lattice or Cayley tree. Near this percolation threshold the critical



exponents with which several quantities diverge or vanish are the same as in the random graphs of Erdős and Rényi. But this percolation theory was published nearly two decades earlier, in 1941 by P. Flory, later to become a chemistry Nobel laureate.

### Mean Field Approximations

What is called “mean field” is similar to the “representative agent” theory in economics, and is widespread in chemistry where the changes in the concentrations of various reacting compounds are approximated as functions of these time-dependent concentrations. A particularly simple example is Verhulst’s logistic equation  $dx/dt = ax(1 - x)$ , known as Bass diffusion in economics. We now explain why this approximation is unreliable.

Let us return to the above-mentioned Ising model of Eq. (2) and replace the  $S_k$  in that equation by its average  $\langle S_k \rangle = m = M/L^d$  which is a real number between  $-1$  and  $+1$  instead of being just  $-1$  or  $+1$ ;  $m$  is the normalized magnetization. Then the total energy  $E$  is approximated as the sum over single energies  $E_i$ :

$$E = \sum_i E_i, \quad E_i = \left( - \sum_k \langle S_k \rangle - B \right) S_i = -B' S_i$$

with a mean magnetic field  $B' = B + \sum_k \langle S_k \rangle = B + mz$  where  $z$  again is the number of lattice neighbors. The system now behaves as if each spin  $S_i$  is in an effective field  $B'$  influenced only by the average magnetization  $m$  and no longer directly by its neighbors  $S_k$ . The two possible orientations of  $S_i$  have the energies  $\pm B'$ , giving an average

$$m = \langle S_i \rangle = \tanh(B'/T) = \tanh[(B + zm)/T] \quad (3)$$

and thus a self-consistency equation for  $m$ . Expanding the hyperbolic tangent into a Taylor series for small  $m$  and  $B$  we get

$$B = (1 - z/T)m + m^3/3 + \dots \quad (4)$$

which gives a Curie temperature  $T_c = z$ , since for  $T < T_c$  the magnetization is  $m = \pm[3(z/T - 1)]^{1/2} \propto (T_c - T)^{1/2}$ . Similar approximations for liquid–vapour equilibria lead to the Van der Waals equation of 1872, which may be regarded as the first quantitative theory of a complex phenomenon. ( $m$  in that case is the difference between the liquid and the vapour density.) Nowhere in Eqs. (2) and (3) have we put in that there is a phase transition to ferromagnetism; it just arises from the very simple interaction energy  $S_i S_k$  between neighboring spins, and similarly the

formation of raindrops emerges from the interaction between the molecules of water vapour. The water molecule is the same  $H_2O$  in the vapour, the liquid or the ice phase.

But this nice approximation contradicts the results mentioned above. For the chain, square and simple cubic lattice it predicts  $T_c = z = 2, 4$  and  $6$  while the correct values are  $0, 2.2$  and  $4.5$ . Particularly in one dimension it predicts a phase transition at a positive  $T_c$  while no such transition is possible:  $T_c = 0$ . This was the main result of Ernst Ising’s thesis in 1925. And even in three dimensions, where the difference in  $T_c$  between  $4.5$  and  $6$  is less drastic, the above square-root law for  $m$  is wrong, since  $m$  varies for  $T$  slightly below  $T_c$  roughly as  $(T_c - T)^{0.32}$ . Thus, mean field theory, Van der Waals equation, and similar approximations averaging over many particles are at best qualitatively correct. They become exact when each particle interacts equally with all other particles.

Analogously for percolation, Flory’s approximation of neglecting cyclic links and the Erdős–Rényi random graphs lead to results corresponding to mean field approximations and should not be relied upon in one, two or three dimensions with links between nearest neighbors only.

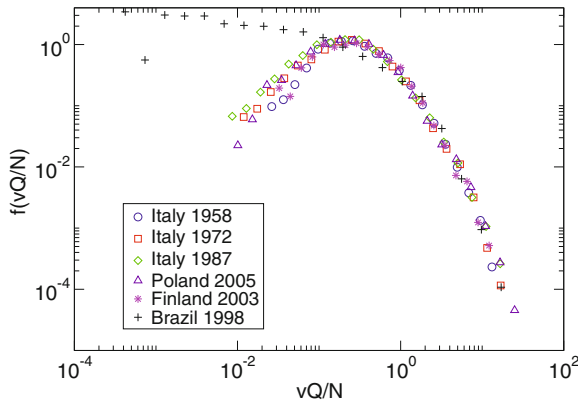
For cellular automata a particularly drastic failure of analogous mean field approximations (differential equations) was given by Shnerb et al. [16] for a biological problem. Even simpler, many cellular automata on the square lattice lead to blinking pairs of next-nearest neighbors: at even times one site of the pair is  $1$  and the other is  $0$ , while at odd times the first is  $0$  and the second is  $1$ . Averaging over many sites destroys these local correlations which keep the blinking pair alive.

### Applications

A thorough review of “sociophysics” was given recently by Castellano et al. [17], and a long list of references by Carbone et al. [18]. In this Encyclopedia some work by social scientists is reviewed by Davidsson and Verhagen in the section on agent-based simulations in sociology, while Troitzsch in this section reviews both social scientists and physicists. His book with Gilbert [29] is, of course, more complete. Thus, we merely sketch here some of the areas covered in greater detail in the other articles or in the cited literature.

### Elections

A social scientist may be interested in predicting the fate of one particular party or candidate in one particular election, or to explain it after this election. A physicist, accustomed to electrons, hydrogen atoms and water molecules



**Physics and Mathematics Applications in Social Science, Figure 1**  
The vote distribution in several countries and elections is a function only of the scaled variable  $vQ/N$ . From [19]

being the same all over the world may be more interested to find which universal properties all elections have in common. Figure 1, kindly provided by Santo Fortunato, is an example. Let  $v$  be the number of votes that a candidate obtained,  $Q$  the number of candidates in that election, and  $N$  the total number of votes cast. Then the probability distribution  $P(v, Q, N)$  for the number of votes is actually a function  $f(vQ/N)$  of only one scaled variable, and that variable  $vQ/N$  is the ratio of the actual number  $v$  of votes per candidate to the average number  $N/Q$  of votes per candidate. Various countries and various elections, all using a proportional election system, gave the same curve  $f(vQ/N)$  which is a parabola on this double-logarithmic plot and thus corresponds to a log-normal distribution. In Brazil, however, where the personality of a candidate plays a major role, with the party affiliation of the candidate playing a lesser role, the results were different. These authors also present a model to explain the log-normal distribution [19].

Other models for opinion dynamics are reviewed elsewhere in this Encyclopedia ► [Opinion Dynamics and Sociophysics](#) or in [5]. A more cognitive approach for interacting agents is their realization by neural network models [20,21,22], ► [Social Cognitive Complexity](#).

### Financial Markets

Agent-based simulation of stock markets [23] are a typical example of complex systems applications: In these models not the single agent but their (unconscious) cooperation produces the ups and downs on the stock market, the bubbles and the crashes. These models deal with the more or less random fluctuations, not with well founded market

changes due to new inventions or major natural catastrophes.

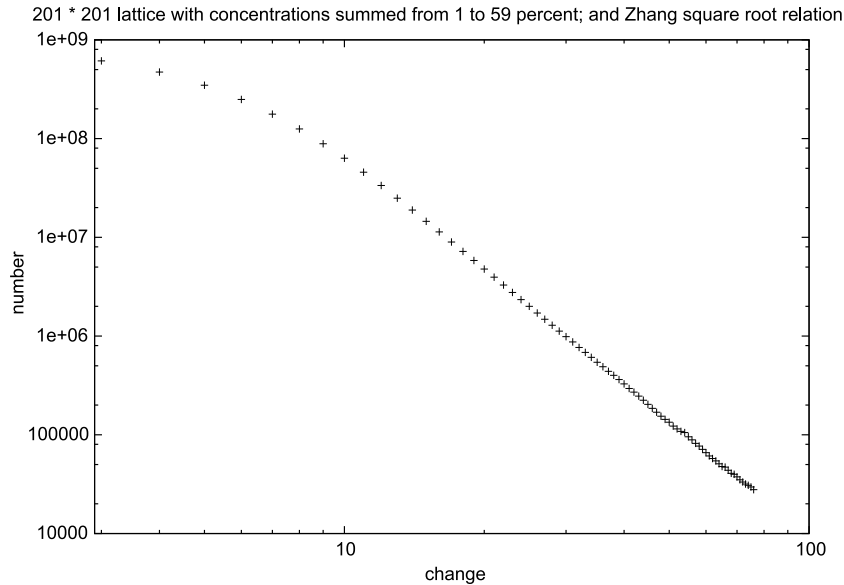
Real markets give at each time interval a return  $r$  which is the relative change of the price. Typically, an index of the whole market like Dow Jones changes each trading day by about one percent. Much larger fluctuations are rarer, and the probability to have a change larger than  $r$  decays for large  $r$  as  $1/r^3$ : Fat tails, compared with normal Gaussian distributions. The sign of the change is barely predictable, but its absolute value is: Volatility clustering. Thus, in calm times when  $|r|$  was small, tomorrow's  $|r|$  probably is also small, whereas for turbulent times with high  $|r|$  in the past one should also expect a large  $|r|$  tomorrow. The daily weather behaves similarly: presumably tomorrow will be like today.

A simple model, going back to Bachelier more than a century ago, would throw a coin to determine whether the market tomorrow will go up or down. This simple random-walk or diffusion model was shown by Mandelbrot in the 1960s not to describe a real market; it lacks fat tails and volatility clustering but may be good for monthly changes. Many better agent-based models have been invented during the last decade and reproduce these real properties; the Cont–Bouchaud model is based on the above percolation theory (see Fig. 2) [24], while the Minority Game is based on the hypothesis that it is better not to be with the big crowd [25].

### Languages

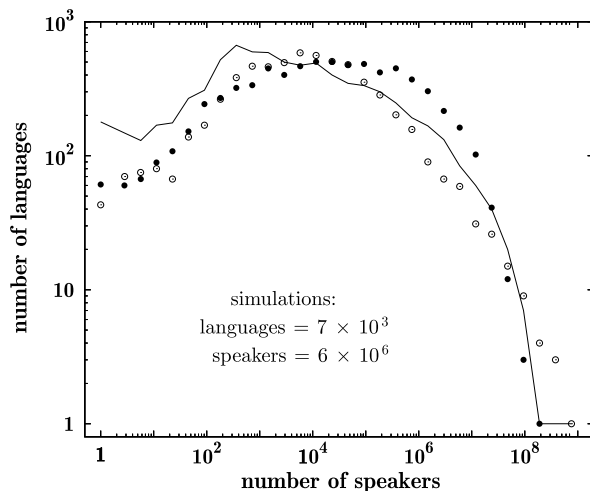
The versatility of human languages distinguishes us from the simpler communication systems of other living beings. With computers or mathematically exact solutions [26] models have been studied for the learning of a language by children or for the evolution of human languages out of simpler forms.

Closer to simulations in biology with the Darwinian selection of the fittest are the models of competition between various languages of adult humans: Will the Welsh language survive against English in Great Britain? Similar to Lotka–Volterra equations for prey and predator in biology, some nonlinear differential equations [12] seem to describe the extinction of the weaker language. Statistics better than in [12] are available for the size distribution of languages, where “size” is the number of people speaking this language, and for the number of languages in one language family. Here one model from de Oliveira et al. [28] found good agreement with reality, see Fig. 3; other models [27] were less successful, in spite of many simulations from physicists.



Physics and Mathematics Applications in Social Science, Figure 2

Simulated return distribution in the Cont-Bouchaud percolation model of stock markets [24]. The asymptotic slope to the right is about  $-2.9$ . We plot the number of simulated events for a given absolute value of the market change (in arbitrary units) versus that value



Physics and Mathematics Applications in Social Science, Figure 3  
Simulated size distributions for human languages (full circles, and line), compared with reality (open circles). From [28]

### Future Directions

The future should see more work on what we have shown here through our three figures: Searching for universal properties, or the lack of them, in the multitudes of models and in reality. Biology became a real science when the various living beings were classified into horses, mammals, vertebrates etc. Within each such taxonomic set all an-

imals have certain things in common, which animals in other taxonomic sets do not share. This check for universality is different from improving our ability to ride horses. Thus, making money on the stock market, or explaining the crash of 1987, is nice, but investigating the exponents of the fat tails, Fig. 2, of all markets may give us more insight into what drives a market and what differences exist between different markets. Winning one particular election and predicting the winner is important, but universal scaling properties as in Fig. 1 may help us to understand democracy better. Preventing the extinction of the French language in Canada is important for the people there, but explaining the overall statistics of languages in Fig. 3 is relevant globally.

It is in these general aspects where the methods of mathematics and physics seem to be most fruitful. One specific problem is better solved by the local people who know that problem best, not by general simplified models.

We thank G. Weisbuch for comments on this manuscript.

### Bibliography

#### Primary literature

1. Schelling TC (1971) J Math Sociol 1:143
2. Dethlefsen E, Moody C (1982) Byte 7:178; Jones FL (1985) Aust N Z J Sociol 21:431

3. Stigler GJ (1964) *J Bus* 37:117
4. Kim GW, Markowitz HM (1989) *J Portf Manag* 16:45
5. Stauffer D (2008) Opinion Dynamics and Sociophysics. *arXiv:0705.0891*
6. Mantegna RN (2005) Presentation of the English translation of Ettore Majorana's paper: The value of statistical laws in physics and social sciences. *Quant Finance* 5:133
7. Weidlich W (2000) *Sociodynamics; A Systematic Approach to Mathematical Modelling in the Social Sciences*. Harwood Academic Publishers, Amsterdam; reprint: Dover, New York
8. Galam S (2008) *Int J Mod Phys C* 19:409; Galam S, Gefen Y, Shapir Y (1982) *J Math Sociol* 9:1
9. Roehner B, Wiese KE (1982) *Environ Plan A* 14:1449
10. Gomes MAF, Vasconcelos GL, Tang IJ, Tang IR (1999) *Physica A* 271:489
11. Zanette D (2001) *Adv Complex Syst* 4:281
12. Abrams MH, Strogatz SH (2003) *Nature* 424:900
13. [http://nobelprize.org/nobel\\_prizes/chemistry/laureates/1921/soddy-bio.html](http://nobelprize.org/nobel_prizes/chemistry/laureates/1921/soddy-bio.html)
14. Stauffer D (1991) *J Phys A* 24:909
15. Ódor G (2008) *Int J Mod Phys C* 19:393
16. Shnerb NM, Louzoun Y, Bettelheim E, Solomon S (2000) *Proc Natl Acad Sci USA* 97:10322
17. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. *Rev Mod Phys* (to be published). *arXiv:0710.3256*
18. Carbone A, Kaniadakis G, Scarfone AM (2007) *Eur Phys J B* 57:121
19. Fortunato S, Castellano C (2007) *Phys Rev Lett* 99:138701
20. Weisbuch G, Duchateau-Nguyen G (1998) *J Artif Soc Soc Simul* 1(2), paper 2 (electronic only: [jasss.soc.surrey.ac.uk](http://jasss.soc.surrey.ac.uk))
21. Wischmann S, Hulse M, Knabe JF, Pasemann F (2006) *Adapt Behav* 14:117
22. Stauffer D (2007) Opinion Dynamics with Hopfield Neural Networks. *arXiv:0712.4364*
23. Levy M, Levy H, Solomon S (2000) *Microscopic Simulation of Financial Markets*. Academic Press, New York; Samanidou E, Zschichang R, Stauffer D, Lux T (2007) *Rep Progr Phys* 70:404; Lux T (2007) Stochastic behavioral asset pricing models and the stylized facts. In: Hens T, Schenk-Hoppé K (eds) *Handbook on Financial Economics*. (draft)
24. Stauffer D (2001) *Adv Complex Syst* 4:19
25. Challet D, Marsili M, Zhang YC (2004) *Minority Games*. Oxford University Press, Oxford
26. Komarova NL (2004) *J Theor Biol* 230:227
27. Schulze S, Stauffer D, Wichmann S (2008) *Commun Comput Phys* 3:271
28. de Oliveira PMC, Stauffer D, Wichmann S, de Oliveira MS (2008) A computer simulation of language families. *J Linguist* 44:659; *arXiv:0709.0868*
29. Gilbert N, Troitzsch KG (2005) *Simulation for the Social Scientist*, 2nd edn. Open University Press, Maidenhead

### Books and Reviews

- Stauffer D, de Oliveira SM, de Oliveira PMC, Sá Martins JS (2006) *Biology, Sociology, Geology by Computational Physicists*. Elsevier, Amsterdam
- Billari FC, Fent T, Prskawetz A, Scheffran J (2006) *Agent-based computational modelling*. Physica, Heidelberg

## Polymer Physics

T. C. B. MCLEISH

UK Polymer IRC, Department of Physics and Astronomy, University of Leeds, Leeds, UK

### Article Outline

Glossary

Definition of the Subject

Introduction

Single Polymer Chain Physics

Equilibrium Properties of Many-Chain Fluids

Dynamics of Polymeric Fluids

Multi-Phase Polymeric Fluids

Future Directions

Bibliography

### Glossary

- $R$**  Radius of gyration of a polymer chain
- $N, M$**  Degree of polymerization or number of monomers in a polymer chain, molecular weight.
- $\xi$**  The screening length or correlation length in semi-dilute polymer solutions; the length over which local density is dominated by a single chain.
- $N_e, M_e$**  Entanglement degree of polymerization and molecular weight.
- $l_p$**  Persistence length of polymer chain.
- $\nu$**  "Flory" exponent of a polymer chain relating  $R$  and  $N$  so that  $R \sim N^\nu$ .
- $S(k)$**  Scattering structure factor from a polymeric fluid as function of scattering vector  $k = 4\pi \sin \theta / \lambda$  where  $\lambda$  is the wavelength and  $\theta$  the scattering angle of the experiment.
- $\Pi(c)$**  Osmotic pressure of a solution as a function of concentration  $c$ .
- $\sigma_{ij}$**  Components of the stress tensor.
- $d, D$**  Dimensions of a macromolecular object and its embedding space.
- $R(n, t)$**  Functional description of a macromolecular contour as functions of monomer number  $n$  and time  $t$ .
- $G(t)$**  Time dependent relaxation modulus.
- $\eta$**  Viscosity.
- $k_B$**  Boltzmann's constant.
- $\chi$**  Flory interaction parameter between monomers of different chemistry.

### Definition of the Subject

Physics is uniquely endowed among the sciences with complete freedom from restriction to any particular do-

main of the physical world. It is able to turn its particular outlook on the scientific program and its special set of experimental and theoretical tools to most material phenomena. In particular it is not limited to any particular length scale, but is sensitive to the emergence of new structures and processes of any size. So macromolecular science, born of the chemistry of the early 20th century, soon gave rise to a branch of physics that seeks to understand the special phenomena of polymer molecules and polymeric matter. Polymers are giant, usually linear molecules constructed as covalently-bonded chains of identical units, or monomers. Individual polymer molecules may contain hundreds or even millions of monomers. Initially disfavored by an organic chemistry community that prized exactitude, polymers were largely ignored since their molecular weight was inexact within a single sample. However, biology knew long before we did that the properties of polymers met the essential requirements of living organisms. Polymers constitute nature's scaffolds from the macroscopic (bone, collagen) to the microscopic (filamentous actin, polymerized tubulin), her force-generators (myosin, dynein and kinesin protein polymers) her information-processing networks (peptide-binding proteins, polysaccharides) and her instruction sets (the nucleic acid family of polymers including DNA). Much of the reason for this lies in the unique set of physical properties of polymeric matter. Some of these have now familiar application in plastic materials from packaging to high-performance fiber and even addressable polymeric electronics. Yet polymer physics is much more than the application of statistical mechanics and spectroscopy to a class of molecular matter; it has taught our discipline about some of its own deep structural connections. Path-integral techniques at the heart of theoretical polymer physics were adopted from readily developed tools in quantum field theory and magnetism. Flexible linear-like objects occur in many other avenues of the subject, superconductivity and plasma turbulence to name two. Above all polymer physics has provided us with a classic example of emergent simplicity from bewildering complexity, so beloved of our subject. It shows every indication of providing in future an essential Ariadne's thread to guide us in our exploration of complexity itself.

## Introduction

The fascinating physics of flexible polymers flows from both necessity and beauty. Born of the early investigations into the phenomenon of the elasticity of natural rubber [1,2], then out of the rapid growth in synthetic polymer materials in the post-war years, the need to under-

stand and control the processing of such highly viscoelastic liquids as polymer melts, and to understand the properties of the resulting materials led rapidly to fundamental investigations led by physicists who saw an opportunity to explore the fundamental structures of a new class of matter. Flory [3], Zimm and Stockmayer [4] and Edwards [5] asked how large would macromolecules, linear or branched, be, while Zimm [6] and Rouse [7] asked how such giant molecules would move. Paralleling developments in solid-state many-body physics, the focus of investigations moved from single-chain to many-chain systems, which we review below in Sects. "Single Polymer Chain Physics" and "Equilibrium Properties of Many-Chain Fluids". These pioneers were already using a beautiful notion that was to take hold of condensed-matter physics in the mid 20th century – that of *universality*, or the independence of physical phenomena from local, small-scale details. The emergence of universal properties is usually associated with "critical phenomena" [8], since near phase transitions, the spatial scale of correlated fluctuations may hugely exceed molecular dimensions. Any properties that depend on these fluctuations (an example would be compressibility of a fluid near its critical point, and especially the *exponent* with which it vanishes as the temperature tends to its critical value) will then be insensitive to molecular detail. In field theories of both condensed and high-energy matter, the field-fluctuations "renormalize" microscopic constants into new emergent numbers on which the physics at coarser length scales (or lower energies) may be built (a famous example is the charge of the electron). Although there is at first glance no apparent neighboring critical point in the case of polymeric fluids, both universality in exponents and renormalized quantities appear in abundance. Moreover, there is a natural large number associated with mesoscopic, rather than microscopic lengthscales. The defining feature of a polymer is, after all, its large "degree of polymerization",  $N$ , the number of monomers linked together covalently to form the polymer chain. (The literature discusses interchangeably  $N$  and the *molecular weight*  $M$  of the chains, given in terms of the monomer molecular weight  $m_0$  by  $M = Nm_0$ ). At the most basic level of inquiry into polymer structure, experiments and simulations asking how the average end-to-end distance  $R$  of a polymer molecule in solution depends on its degree of polymerization  $N$ , began to suggest a universal scaling behavior

$$R \sim N^\nu \quad (1)$$

with an exponent  $\nu$ , dependent only on the embedding dimension  $d$  and first calculated by Edwards to be  $\nu = \frac{d}{d+2}$ . It assumes a rather larger value in solution ( $\simeq 0.59$ ) than



the simple random walk value of 0.5 [9], due to the self-exclusion of the monomers. More phenomena reminiscent of other areas of condensed matter appeared at the level of many-body effects. In the dense limit of polymer melts and concentrated solutions, where chains are highly overlapped (and in embedding dimensions greater than 4, the “upper critical dimension” of the self-exclusion problem), the exponent  $\nu$  reassumes the value of 1/2 of the ideal Gaussian random walk (“Gaussian” because the ensemble of spatial end-to-end vectors of the polymer chains is normally distributed). Closer inspection revealed this to be true above a “screening length”, introduced into polymer physics by Edwards [10]. The screening length  $\xi$  itself may be directly measured by neutron scattering, and depends on concentration via another universal scaling exponent, related to  $\nu$  [11]

$$\xi \sim c^{\frac{-\nu}{3\nu-1}}. \quad (2)$$

The picture we have build up of a many-chain polymer solution so far is summarized in Fig. 1, where atomic detail at the monomer level is far below the resolution of the diagram. As the polymer concentration increases, so the screening length or “mesh size” decreases. A typical strand of chain, whose end-to-end distance is  $\xi$ , dominates the monomer concentration within the volume it spans.

Both experimental and theoretical evidence of universality continued to build up. Even in the case of dynamics, the many-chain system of a polymer melt followed the ideal, local-dissipation theory of Rouse [7] for sufficiently low molecular weight chains (see below Sect. “Dynamics of Polymeric Fluids”) that assumed ideal Gaussian chains. It became clear that Rouse’s result can be seen as a fixed

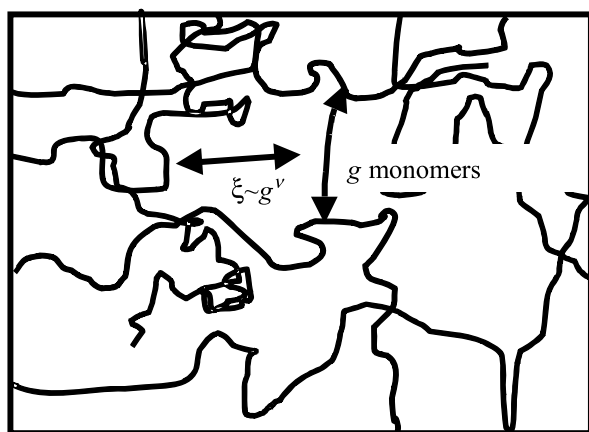
“point” of all theories of polymer dynamics in which linear connected objects are ideal and are subject to local dissipation (in dilute solution, far-field hydrodynamics destroys this locality [6]). For example, lattice models of polymer dynamics with local update rules renormalize to the continuum Rouse theory at large enough length scales [12].

It seemed as though the huge connectivity of macromolecules acts to freeze-in long-range order, even though there is no true thermodynamic transition nearby. Such suspicions were confirmed by the demonstration of direct isomorphisms of the calculation of statistical mechanical partition functions of polymers, both dilute and concentrated, onto idealized spin-lattice models of magnetism [9]. It is indeed the high molecular connectivity, as the inverse of the degree of polymerization,  $N^{-1}$ , that plays the part of proximity to the distance from a critical point in the spin model

$$N^{-1} \sim \varepsilon \equiv \frac{T - T_c}{T_c}. \quad (3)$$

So by exhibiting physics in which an ensemble of macromolecules of polystyrene (PS) exhibits the same emergent behavior as polyisoprene (PI) or polybutadiene (PB), following scaling laws, and tractable by application of statistical mechanical field theories [8], polymer physics drew together many of the strongest conceptual strands of the century.

More, however, has proved to be true in the realm of *topological* effects. The polymer melts of industrial polymer processing are very highly overlapped on the molecular level, where it becomes immediately apparent that molecular relaxation processes controlling elastic stress are prolonged to very long times indeed. All the important phenomenology is covered in Ferry’s seminal survey of polymer viscoelasticity [13]. Mechanical experiments restricted to a range of intermediate timescales of the plateau are hardly able to distinguish between the polymer melt and a rubber, in which the chains are permanently cross-linked to each other at very rare points, sufficiently for each chain to be permanently immobilized from large-scale diffusion. Conceptually, the absent “cross-links” were replaced in the minds of engineers and physicists alike by “entanglements” [13]. These loosely-defined objects were assumed to represent the topological constraint that covalently-bonded molecular chains may not pass through each other. The effective distance between these objects could be calculated, employing rubber elasticity theory (see below), to deduce the degree of polymerization between entanglements  $N_e$ , or the equivalent “entanglement molecular weight”,  $M_e$ . The number



**Polymer Physics, Figure 1**

Schematic picture of universal structures of screening (overlap) length  $\xi$  and the number of monomers  $g$  that just spans  $\xi$

$N_e$  consistently turned out to be of order  $10^2$ , indicating a length-scale for an “entanglement spacing” of 50–100 Å, depending on the particular chemistry. This is highly significant for us, because it shows that small chains on the threshold of feeling topological interactions are real polymers, already long enough to show to a good approximation all the universal properties of statistical connected chains. It also suggests that the role of topology in highly-entangled ( $N \gg N_e$ ) polymer fluids has the potential to be treated universally. Further evidence of universality in entanglements came from experiments in which the polymers were diluted to a volume fraction  $\phi_p$  by a compatible solvent, indicating that  $M_e \sim \phi_p^{-\alpha}$  where the scaling exponent  $\alpha \simeq 1$  [11]. The dependence of melt viscosity  $\eta$  (at fixed temperature) on molecular weight also exhibits remarkable universality over very many different polymer chemistries [13] closely matched by  $\eta \sim M^{3.4}$ , providing that the molecular weight lay well above the entanglement threshold  $M_e$  suggested by their high-frequency elastic modulus. We review this rapidly-progressing area in Sect. “Dynamics of Polymeric Fluids”.

New phenomena arise when different chemistries of monomer are introduced into the same chain. Effective repulsive interactions between heterogeneous monomers create a tendency for strong spatial correlations of chemical type. In blends of more than one type of homogeneous polymer, the result is often a demixing transition with near-universal structure and dynamics [17]. When the chemical species are combined into the same chain in regular “blocks” of controlled molecular weight, demixing occurs on the scale of the chains themselves, giving an extremely rich variety of spatially-periodic nanoscopic structures, self-assembled micellar structures, and controlled collapsed forms. Both chain composition and temperature act as control parameters of a structural space that becomes increasingly biomimetic as the information content of the macromolecular sequence increases. Experiment and theory are reviewed in Sect. “Multi-Phase Polymeric Fluids”.

### Single Polymer Chain Physics

It should not be surprising that the notion of complexity should arise even in the context of the “single particle” domain of polymer physics: that of a single molecule. For already a macromolecule contains many degrees of freedom that are coupled in non-linear ways. Not only this, but also at the single-chain level emergent co-operative properties arise. We briefly survey three important cases of single-chain physics: the non-interacting chain, the effect of excluded volume and the role of charge.

### Ideal Non-interacting Chains

The first and most fundamental of these underlies the physics of rubber elasticity: it is the emergence of an entropic Hookean spring for macrostates of a single polymer chain in thermal equilibrium defined in terms of its end-to-end vector  $\mathbf{R}$ . We recap briefly here the statistical physics of a polymer chain, modeled as a *random walk* in space and subject to some local rule for spatial links. An example is the freely jointed chain, in which the orientations of a set of linked rods are uncorrelated. The step length of the chain corresponds to the Kuhn length,  $b$ , of the polymer (the shortest independently oriented segment length). It is not as small as a monomer length, but usually 4 or 5 monomers long.

For polymer statistics suppose a whole walk has  $N$  links. Let the end to end displacement of an individual chain be  $\mathbf{R}(N)$ . From the theory of random walks:  $\langle R^2(N) \rangle = Nb^2$  and the probability density for the end to end vector  $G(\mathbf{R})$  must have Gaussian form (from the law of large numbers, since each vector step is an independent random variable whose sum is  $\mathbf{R}(N)$ ). So

$$G(\mathbf{R}, N) = \left( \frac{3}{2\pi Nb^2} \right)^{3/2} e^{-3R^2/2Nb^2}. \quad (4)$$

The macrostate of an ensemble of such chains is defined by the chain end-to-end vector  $\mathbf{R}$ . The microstates are the different specific paths through space that have  $\mathbf{R}$  as their end-to-end displacement. Each individual path, or microstate of the chain, will be specified if the spatial position of each link is known. We will use the notation  $\mathbf{R}(n)$  for the position of the  $n^{\text{th}}$  link. The full time-dependence of the chain would then be described by the function  $\mathbf{R}(n, t)$ , extended to the two dependent variables of contour position  $n$  and time  $t$ . The role of Brownian motion can be cast in the form of Langevin equations for  $\mathbf{R}(n, t)$  (see Rouse model, Sect. “Dynamics of Polymeric Fluids”), but here we exploit it as a generator of ergodic exploration of all chain configurations in the ensemble. The number of configurations with fixed end to end vector  $\mathbf{R}$  is just the corresponding fraction of total microstates  $\Omega(\mathbf{R}) = \Omega_{\text{TOT}} P(\mathbf{R})$ . Since the entropy of the walk  $S = k_B \ln \Omega(\mathbf{R})$  we have  $S(\mathbf{R}) = \text{const.} - \frac{3k_B R^2}{2Nb^2}$ . The conformational free energy of the chain  $F(\mathbf{R}) = U - TS$  has  $U = 0$  since there are no sources of internal energy. This yields for the free energy of a chain of fixed end-to-end vector  $F(\mathbf{R}) = \frac{3k_B T R^2}{2Nb^2}$ . Finally we may derive the thermodynamic force (or “Brownian tension”) on the chain end-to-end vector as

$$f = -\nabla F(\mathbf{R}) = -\frac{3k_B T}{Nb^2} \mathbf{R} \quad (5)$$

and recognize a linear elastic spring law. I. e., a random walking polymer at finite  $T$  is a Hookean spring with spring constant  $\propto T/N$ .

The analogies between the statistical mechanics of random walks and the quantum mechanics of spinless particles also emerges from this analysis. Applying Eq. (4) to small subchains of path length  $\Delta N$  allows the (careful) taking of a limit so that

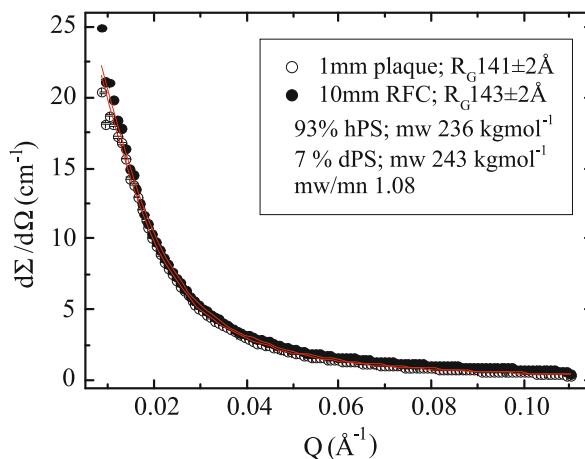
$$G(\mathbf{R}, \Delta N) = \left( \frac{3}{2\pi\Delta N b^2} \right)^{3/2} e^{-\frac{3}{2b^2} \left( \frac{\partial \mathbf{R}}{\partial N} \right)^2 \Delta N}.$$

The notation  $G$  for the probability distribution is suggestive: this descriptor of the chain has the structure of a propagator. The complete end-to-end propagation of (4) can be written as the sum over all possible intermediate positions of the meeting points of all smaller subchains, which in turn is just an example of a Feynman path integral over all possible paths of the polymer contour  $\mathbf{R}(n)$

$$G(\mathbf{R}, N) = \int_{\mathbf{R}(0)=0}^{\mathbf{R}(N)=\mathbf{R}} e^{-\frac{3}{2b^2} \int_0^N \left( \frac{\partial \mathbf{R}}{\partial n} \right)^2 dn} \mathcal{D}[\mathbf{R}(n)]. \quad (6)$$

The propagator structure arises because the properties of the chain at equilibrium is governed by its partition function, which in turn is a sum over microstates that are in this case just the possible paths of the chain. An analogous integral arises in Feynman's form of quantum mechanics because the sum over all trajectories that gives the (complex) amplitude for particle propagation is just the same geometrical set of paths. The difference is that in polymer statistical mechanics the phase angle attributed to the path is imaginary, giving a real argument of the exponential in the path integral. A rich set of techniques flow naturally from this analogy: since the propagator also obeys Schrödinger's equation, the equilibrium configuration of ideal chains in confined geometries and external potentials can be solved by any technique developed for the quantum mechanical case [9].

Experiments on single chains have until recently been indirect ensemble measurements. However, neutron-scattering can give averaged single chain properties because of the very different scattering lengths of hydrogen and deuterium nuclei. It is relatively straightforward to replace chemically some or all of the hydrogen atoms in a fraction of the chains in a polymeric fluid. When the chains themselves are monodisperse, the small-angle scattering pattern is identical to that of the population of labeled chains. Figure 2 shows an example of scattering from a 7% labeled fraction of polystyrene chains with a narrow distribution of molecular weight. Early experiments of this type showed the remarkable result illustrated here that the

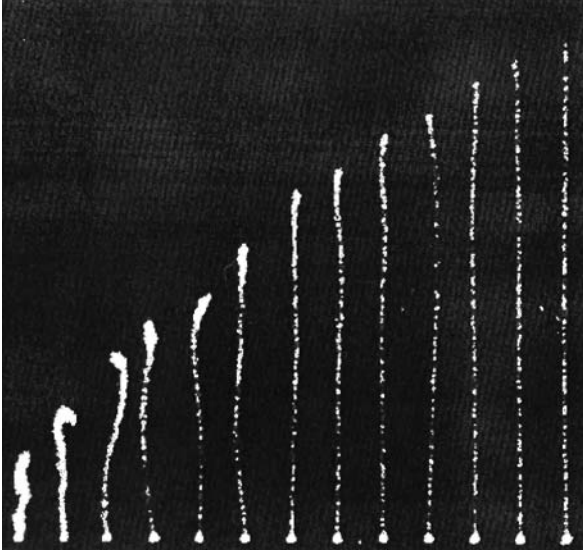


**Polymer Physics, Figure 2**

Small angle neutron scattering (SANS) data averaged over angles for a monodisperse polystyrene melt. The theoretical curve is that calculated from a Gaussian propagator

single chains in a densely packed melt actually assume an ideal (effectively non-self-interacting) set of configurations described by the Gaussian propagator of Eq. (4) (see Sect. “Equilibrium Properties of Many-Chain Fluids” below). Much more recently the advent of recombinant DNA, fluorescent labeling and video confocal microscopy has begun to make direct inspection of individual polymer molecules possible in restricted circumstances. Even very high molecular weight DNA has a random-walk molecular dimension ( $bN^{1/2}$ ) below the resolution limit of optics, but if the chain is stretched out much larger dimensions are accessible approaching  $bN$ , so that some of the predictions of single-chain elasticity can be explored.

Figure 3 shows one famous example that actually illustrates chain response as the maximum elongation is approached. In this limit the Gaussian approximation breaks down badly and a divergence of force with extension is measured. The form of the divergence is, unlike the linear entropy-dominated range of elasticity, not universal among local polymer chemistries, but depends on the form of the local structure and its response to tension. Another analogy with high-energy physics arises here: *strong* applied forces measure structure at *smaller* length scales. This is powerfully illustrated by the high-force asymptote of two models of flexible polymers, the Freely Jointed Chain (FJC) and the Worm-Like Chain (WLC). The first models the local structure of a chain as  $N$  freely-hinged but infinitely stiff rods each of length  $b$ . Such a chain has a maximum extension of its own contour length  $L_0 = Nb$ , but at low forces responds with the linear behavior of the Gaussian chain so that the force  $f$  with extension  $L$  follows



**Polymer Physics, Figure 3**

Images of a  $64\mu$  long DNA molecule held in place at one end by optical tweezers and stretched out by hydrodynamic flow of increasing velocity from left to right. Reprinted with permission from [19]

$f \sim k_B T / (L N b^2)$ . The force diverges as  $L \rightarrow L_0$  with the asymptotic form  $f(L) \sim (1 - L/L_0)^{-1}$ . The WLC introduces a finite bending rigidity everywhere along the chain so that the internal energy of a configuration  $\mathbf{R}(n)$  can be written

$$E = k_B T l_p \int_0^N \left( \frac{\partial^2 \mathbf{R}(n)}{\partial n^2} \right)^2 dn \quad (7)$$

with the constraint that  $|\frac{\partial \mathbf{R}(n)}{\partial n}| = 1$ . The stiffness is written as  $k_B T l_p$  in terms of the “persistence length”  $l_p$  because at equilibrium the chain is locally stiff at smaller length-scales and flexible over longer lengths. This statement can be made exact by considering the correlation function of the orientation of the chain:

$$\left\langle \frac{\partial \mathbf{R}(n_1)}{\partial n} \frac{\partial \mathbf{R}(n_2)}{\partial n} \right\rangle = e^{-|n_1 - n_2|/l_p}.$$

Although this model also shares the Gaussian linear response of the FJC, it possesses quite different asymptotics at high force [20], following  $f(L) \sim (1 - L/L_0)^{-1/2}$ . The calculation takes the response under the force of all harmonic normal modes of the Hamiltonian (7), the more gentle divergence arising from the successive suppression of contortions of the chain at smaller and smaller wavelength as the force increases.

### Self-Interacting Chains: Excluded Volume

As pointed out in the introduction, real polymer chains in solution are not typically Gaussian. The reason is that the path integral of (6) overcounts the allowable configurations of the chains, including those that cross themselves. The modification to the Hamiltonian that achieves the monomeric self-exclusion with maximum simplicity and generality is the “Edwards Hamiltonian” [21]

$$G(\mathbf{R}, N) = \int_{\mathbf{R}(0)=0}^{\mathbf{R}(N)=\mathbf{R}} e^{-\frac{3}{2b^2} \int_0^N \left( \frac{\partial \mathbf{R}}{\partial n} \right)^2 dn} e^{-w \int_0^N \int_0^N \delta[\mathbf{R}(n) - \mathbf{R}(n')] dn dn'} \mathcal{D}[\mathbf{R}(n)].$$

Although formally the delta-function potential removes only a volume of phase space of zero measure from the path integral, its anticipated use within field-theoretic tools for the solution of the model mean that it represents a renormalized local repulsion between monomers of the chain at the level of this universal coarse-grained theory. From this starting point one can proceed by several methods: [21] self-consistent field treatment of the excluded volume term, mapping onto problems in critical phenomena [23], direct renormalization-group calculation [24] and Monte Carlo numerical enumeration [22]. For a comprehensive and technical review see [25]. The essential structure within the self-consistent field methods is the same as that of an early calculation by Flory [3] who balanced the scaling forms of the free-energy contributions from chain elasticity ( $R^2/N$ ) and excluded volume ( $N^2/R^d$ ) in  $d$ -dimensional space and minimized to give the dependence of the scaling exponent for chain size  $\nu$  as

$$\nu = \frac{d}{d+2}. \quad (8)$$

This is fortuitously accurate in all dimensions from  $d = 1$  (where it is exact) to  $d = 4$ , which it correctly identifies as the upper critical dimension of the problem since the Gaussian exponent of  $\nu = 1/2$  is recovered here. However, the method is unreliable for the calculation of other quantities because in dimensions less than four the excluded volume term is not perturbative for any strength of the parameter  $w$  in the large- $N$  limit. This really forces a renormalization approach to the problem, already achieved in the case of spin interactions in magnets at the ferromagnetic critical point. A formal and beautiful exact mapping first pointed out by de Gennes exists between the excluded volume chain and an analytic continuation of the Heisenberg model in which the number of spin components goes to zero [23]. The analogy arises because the calculation of



the correlation function between the magnetic spins on any two sites of the lattice in the magnet model can be written as a weighted sum over all non-intersecting paths that connect the two sites within the lattice. So formally

$$\langle S_i S_j \rangle = \sum_N \Omega(N) e^{-\epsilon N}, \quad (9)$$

where  $\Omega(N)$  counts the number of self-excluding walks of length  $N$  connecting the sites and  $\epsilon$  measures the dimensionless temperature difference from the ferromagnetic critical point as in Eq. (3). The analogy led directly to the first calculated values for the Flory Exponent  $\nu$  using diagrammatic expansion methods developed originally from Feynman's perturbation tools for quantum electrodynamics. In three dimensions the problem is non-perturbative – the exponent departs from the mean-field value for all values of the excluded volume parameter  $w$  no matter how small, providing that the chains are long enough. The calculation of  $\nu$  requires first taking the problem in 4 dimensions, where it becomes perturbative, then taking a double expansion in  $w$  and in the analytic continuation of dimension below 4. For short enough segments on the other hand, the chains do not depart strongly from ideal behavior. There is a characteristic length scale at which the energy of self-exclusion equals the thermal energy  $kT$  below which statistics are near-ideal and beyond which the chains are swollen. Subchains of this intermediate length-scale are known as “thermal blobs”. In  $d = 3$  the value of  $\nu \simeq 0.588$ , in close agreement with the value predicted (fortuitously) by (8).

The behavior of chains under an attractive 2-body interaction constitutes an emergent phenomenon that mirrors that under repulsion discussed above. The experimental case is that of a “poor solvent” where monomers of the polymer now enjoy a favorable interaction. Now the chains are densely packed for high enough molecular weights so that  $\nu = 1/d$  above the thermal blob size. In the limit of infinite molecular weight the transition from a swollen to collapsed chain becomes thermodynamic. This is possible to realize experimentally in some cases by simply controlling the solvent quality though control of temperature. There exists in these cases a critical point known as the “theta temperature” at which the effective 2-body monomer-monomer interactions from excluded volume and solvent-induced attraction exactly cancel, leaving only 3-body and higher terms. The “coil-collapse” transition is not sharp (only becoming so in the thermodynamic limit of infinitely long chains).

The collapsed and swollen chain configurations of real chains leave their traces on the emergent elastic properties of single chains that we examined above in the ideal case.

For example, now the effective Hookean spring force-distance relation  $f(R)$  is modified to  $f \sim R^{\frac{\nu}{1-\nu}}$  in general.

### The Role of Electrostatic Charge

A recently very active area in polymer physics that has produced a number of surprises is the form of emergent behavior arising from the combination of electrostatic charge, polymeric connectivity and counter-charges in solution. This is the case of “polyelectrolytes” – polymers containing charged monomers. Complex emergent behavior is perhaps unsurprising since all three have the propensity to generate long-ranged interactions that are candidates for generators of qualitatively different physics from the local interactions examined in the last section.

The static configurations of a polyelectrolyte are radically different from those of a neutral flexible chain in solution. Constructing a mean-field “Flory” type theory for a chain containing a fraction  $f$  of monomers carrying charge  $e$  that, balancing electrostatic energy with configurational entropy yields the result [26]

$$R \simeq Nb f^{2/3} \left( \frac{l_B}{b} \right)^{1/2} \quad (10)$$

showing that at this level the chains will be completely stretched at the scaling level. This result contains one of the several new length-scales that appear in complex fluid electrostatics, the Bjerrum length

$$l_B \equiv e^2 / (\epsilon kT).$$

This is the distance between charges at which the electrostatic and thermal energies are comparable.

Charged configurations of this extreme kind are not realized globally, since counter-ions are universally present to ensure overall neutrality. Their presence in solution screens the “bare” long range electrostatic repulsion beyond the “Debye screening length”

$$l_D \equiv \left( \frac{\epsilon kT}{n_0 e^2} \right)^{1/2}.$$

This is not the length-scale on which the charged chains adopt random-walk configurations, however, since there is typically a strong repulsion (many  $kT$ ) between chain segments adjacent by  $l_D$ , providing the charge fraction is great enough. The emergent persistence length is another new scale

$$l_p = l_D \left( \frac{l_D l_B}{b^2} \right) f^2.$$



A different set of structures arises if the energy competition is between electrostatic repulsion and chain-collapse in a poor solvent. Without electrostatics the chain will collapse into a compact globular form with  $\nu = 1/d$  to minimize the total contact between monomers and solvent, restricting it to the surface of the globule. The surface energy rises as  $\gamma R^2$ , but is eventually overcome by the electrostatic self-repulsion from the globule's charge, which rises (in three dimensions) as  $N^2 f^2 e^2 / R \sim R^5$ . This instability was identified as a cause of droplet breakup in simple fluids by Lord Rayleigh. The free-energy minimum is achieved in the polymeric case by a configuration resembling a string of pearls, in which the polyelectrolyte breaks up into small globules that closely balance the surface and electrostatic energies, connected by stretched strands of chain. It is a rare example of heterogeneous configurations minimizing the free energy at the level of a single chain [27].

The combination of persistent, or rod-like configurations and counter-charges leads to other remarkable properties of charged polymers. The entropy of confinement of counter ions within a cylindrical region around an extended polymer of radius  $R$  decreases as  $kT \log R$  per ion. In this geometry this has the same functional form as the electrostatic attraction (providing that  $R$  is less than the screening length), which depends on the charge per unit length of the polymer  $\rho$  as  $(\rho e/\epsilon) \log R$ . Providing  $(\rho e/\epsilon) > kT$ , most of the counter-ions in solution balancing the charge on the polymer will be closely-bound to it, the phenomenon of "Manning condensation" [28].

The charge clouds condensed in the vicinity of neighboring polymers sustain thermal fluctuations that correlate via the long-range electrostatic interaction in an analogous way to the electronic origin of the Van der Waals interaction. Providing the fluctuations are large enough (these are enhanced by ions of high valency), then the fluctuation-induced attraction between the counter-ion clouds may actually overcome the electrostatic repulsion of the bare polyelectrolyte charge. For polymers in solution like charges may indeed attract! This effect is responsible for the attraction and bundling of polymers of DNA, which carries a strong negative charge [29].

### Equilibrium Properties of Many-Chain Fluids

We have seen that the thermodynamic properties of single polymer chains exhibit both richness and a high degree of universality. Both aspects of the emergent properties of polymer physics extend to many-chain systems in which the molecular chains become strongly overlapped. When this happens the quality of the solvent (controlled with

temperature) that for single chains gave rise to the coil-collapse transition, swollen statistics and the pseudo-ideal theta temperature, now control co-operative phenomena such as the emergence of an osmotic pressure, and phase separation.

### Semi-dilute Solutions

The existence of the coil-size itself  $R \sim N^\nu$  sets a new concentration regime, termed "semi-dilute", that only exists in the case of polymeric fluids. We have already encountered it, and its key attributes, in the introductory discussion of Fig. 1. The semi-dilute concentration range covers those cases where the chains themselves are strongly overlapped but where the volume fraction taken up by monomer rather than solvent molecules is still small. The structure of the solution when the polymer-solvent interaction is favorable is entirely dominated in this regime by the new lengthscale of the screening length  $\xi$ . The physics of this lengthscale is readily detected in two ways: by direct structural probes such as neutron scattering, and by the solution property of osmotic pressure. If scattering contrast exists between monomer and solvent then the scattering intensity and scattering wavevector  $k$  measures the fluctuations in monomer concentration within volumes of fluid of size  $k^{-1}$ . For large lengthscales  $k \ll \xi^{-1}$ , there are no correlations between one element bounded by a screening length and its neighbors: the identity of a single chain is lost at these scales where it is highly overlapped with others. At the other limit, all scattering for  $k \gg \xi^{-1}$  is governed by correlations along single chain segments within such correlation volumes, and the scattering is identical to that from a single chain. Since chains have a spatial scaling structure whether in theta or good solvents, power-law behavior is induced in the scattering as well. The physics is captured by a generalization of the Ornstein-Zernicke scattering function:

$$S(k) \simeq \frac{S(0)}{1 + (k\xi)^{1/\nu}} \quad (11)$$

illustrated in the figure [30]. The exponent  $\nu$  is the same as that determining coil size for single chains discussed in the previous section.

The osmotic pressure as a function of monomer concentration  $\Pi(c)$  of a semi-dilute polymer solution is also connected with the structure of closely-packed correlation volumes that emerges from the screening-length picture. This is because  $\Pi$  measures the change of free energy with concentration, itself dominated by the balance of chain entropy and chain contact-energy. Since all sub-chain con-

figurations within each correlation volume  $\xi^3$  are sampled ergodically, but all correlations at greater lengthscales lost, the consequence is that the free-energy density carries the structure of  $kT/\xi^3$ , or one thermal degree of freedom per correlation volume. Since the concentration of the correlation length is calculable in terms, once more, of the Flory exponent  $\nu$ , from Eq. (2), the concentration dependence of the osmotic pressure follows  $\Pi(c) \sim c^{3\nu/(3\nu-1)}$  in the semi-dilute regime. The same result can be arrived at by anticipating a scaling structure that crosses-over to the correct ideal-gas form under truly dilute conditions, writing

$$\Pi(c) = \frac{kT}{b^3} \frac{c}{N} f\left(\frac{c}{c^*}\right). \quad (12)$$

Here  $f(x)$  is a scaling function that tends to unity for small argument, and to a power law  $f(x) \sim x^z$  for largest. The concentration  $c^*$  is the overlap threshold separating dilute and semidilute regimes, where individual coils just begin to overlap. Insisting that for  $c \gg c^*$  the osmotic pressure should not depend on chain length  $N$  (this is equivalent to the loss of correlations beyond  $\xi$ ) fixes the power  $z = \frac{1}{3\nu-1}$  and the result  $\Pi(c) \sim c^{3\nu/(3\nu-1)}$  is recovered. Experiments at different molecular weights and chemistries in good solvents collapse well onto the scaling form of (12) [31]. A formal route to the calculation of the solution properties of correlation, free energy and osmotic pressure was discovered by Des Cloiseaux as a generalization of the magnetic spin-analogy of de Gennes. A diagrammatic expansion of  $n$ -body chain interactions is performed for the zero-spin limit of the model, but this time in the presence of an external field. This acts as a fugacity for chains, and creates a system where renormalization-group methods may be used, again in expansion around 4 spatial dimensions, to calculate exponents such as  $z$  [25].

**Complex Topology Polymers and Gelation** The emergent properties of polymers in solution, it should be clear by now, arise from the connectivity of chains, either on its own, or in the presence of other physical interactions such as excluded volume or electrostatics. It is also the essentially topological properties of connectivity that endow polymer physics with its universality. It is therefore of interest to explore the consequences of altering the chain topology in more complex ways. One very practical way of doing this is to introduce cross-links chemically into a polymer solution (industrially this is the route to preparation of rubbers). In the limit of high cross-link density the result is a system of chains that have a significant number of their degrees of freedom “quenched” (chemically connected monomers on different chains are perpet-

ually forced into proximity). That the resulting physical object is a solid when the original system was fluid is by no means an obvious result, and emerges only delicately from the treatment of the statistical mechanics of quenched disorder. Historically, methods originally developed to treat the model systems of “spin-glasses” were the first to treat the cross-linked polymeric fluid from this point of view, among them the “replica method” of Edwards [33]. Here the formal free energy of the rubber as an average over the logarithms of quenched partition functions with different configurations of cross-links  $\{N_x\}$ :

$$F(N, N_x, \Lambda) = kT \langle \ln Z(N, \{N_x\}, \Lambda) \rangle_{\{N_x\}}$$

is treated using the formal limit  $\log Z = \lim_{n \rightarrow 0} \left( \frac{Z^n - 1}{n} \right)$ . This amounts to the statistical mechanics of an unquenched ensemble of replicas of the original network, but in the limit (analytically-continued) of the number of replicas tending to zero. The liquid-solid transition emerges as the physical consequence of a mathematical symmetry-breaking between the replicas in the evaluation of the free energy, and a shear-modulus grows as the third power of the difference between the cross-link density and a critical value for the onset of the solid [34].

Although the formal treatment of rubber elasticity is very subtle, there are good approximations that have generated a sequence of semi-empirical models capable of capturing not only stress generation in cross-linked rubbers, but also the temporary stress generated in entangled polymeric fluids (see below). These begin with identifying a length-scale at which the deformation  $\Lambda$  is imposed on chain segments affinely, and below which the chains are assumed to be able to explore all microstates. Applying the result for the effective elasticity of the chain segments (5) above yields an expression for the bulk stress tensor in terms of the average configuration of the subchains:

$$\sigma_{ij} = \frac{3k_B T}{b^2} \mathbb{C} \left\langle \frac{\partial R_i}{\partial n} \frac{\partial R_j}{\partial n} \right\rangle. \quad (13)$$

These results obtain for the case in which the cross-links are suddenly imposed upon an equilibrium semi-dilute or concentrated polymer solution. If they are introduced gradually a richer structure arises in which chains of increasingly branched nature are created well before the critical liquid-solid transition. If this process is imposed mathematically on an ensemble of ideal chains, an ensemble of very compact branched polymers results with a Flory exponent of  $\nu = 1/4$ . From the definition of  $\nu$ , (1), we see that its inverse  $1/\nu \equiv D$  plays the role of a dimension. Ideal linear chains are in this sense two-dimensional ob-

jects, but ideal branched polymers are four-dimensional! This does not present a necessary difficulty at fixed spatial scales, since the individual molecules of finite molecular weight are very sparse, but at high enough molecular weight there will not be sufficient space in three embedding dimensions to contain such objects without overlap. This must occur in successive cross-linking, because calculations of the distribution of molecular weights that result from random insertion of cross-links (this may be done elegantly by the use of generating functions [9]) gives a result of the form

$$P(M) = M^{-\tau} f(M/M_x), \quad (14)$$

where  $\tau$  is a “Fisher exponent” and  $f(x)$  a cut-off function that limits the distribution function to a highest molecular weight of  $M_x$ . This upper molecular weight itself diverges as another scaling function of the difference in cross-link density from the critical value  $M_x \sim |p - p_c|^{-1/\sigma}$ . This class of critical phenomena generated by topological connectivity is called “percolation” and generates beautiful physical effects in polymers. Like other critical phenomena associated with thermodynamic phase transitions it possesses universality classes in which the values of the exponents are independent of the geometry of local interactions. We can see that the mean-field (ideal) value of  $D = 4$  (and  $\tau = 5/2$ ) is not sustainable in 3-dimensional space: the branched polymers must swell by excluded volume so that they do not overlap strongly with themselves (or with subclusters of larger molecules) [35]. Applying this requirement on the non-overlap of clusters of all molecular weights generates a relationship between the exponents  $\tau$ ,  $D$  and the embedding dimension  $d$  called the hyperscaling relation:

$$\frac{d}{D} = \tau - 1. \quad (15)$$

This holds for a wide range of “bare” fractal dimensions  $D$  since the excluded volume drives systems to marginal overlap at all length scales stably: if the overlap reduces for larger chains then they suffer less excluded volume and swell less, consequently increasing overlap once more. The values for percolation in 3-dimensional space are  $D \cong 2.53$  and  $\tau \cong 2.18$ . Should the overlap increase then the consequent increased self-repulsion increases swelling with the opposite effect. The structure implied by (15) is self-similar on a grand scale: not only are the individual molecules self-similar, but the ensemble also satisfies self-similarity at all length-scales in the marginal overlap of clusters of all sizes. There is no correlation length and the scattering function is a power law.

## Dynamics of Polymeric Fluids

Complex, emergent phenomena extend from static structure of polymeric fluids into their rich dynamic properties. This is especially true of the semi-dilute and concentrated cases when interactions between distinct chains are very strong and the conformational dynamics highly coupled. But even in the case of dilute or effectively uncoupled dynamics the scaling structures we saw in equilibrium pattern the behavior out of equilibrium. The experimental tools deployed also mirror those used for statics: direct structural probes of light or neutron scattering possess dynamic counterparts in photon correlation spectroscopy [36] and neutron spin-echo [37]. Specific averages over bond correlation dynamics are available from dielectric spectroscopy [39] or NMR relaxation [38]. Emergent effects at the level of the fluid are most striking in their rheology [40]. At the level of linear response, the rubber-elastic stress generated by small deformations decays with a function  $G(t)$  characteristic of the molecular structures present: the phenomenon of viscoelasticity. In strongly non-linear flows other effects appear, such as the effective decrease in viscosity with shear rate, complex transient behavior in stress response and the generation of rich stress-fields and non-inertial elastic instabilities in non-trivial flows. We first review briefly the dynamic properties of single polymer chains before treating more extensively the complex dynamics of entangled systems.

### Single Chain Dynamics

In the simplest fundamental model of polymer dynamics, due to Rouse, we make three key simplifying assumptions [7]. Their physical validity depends on the effectiveness of screening [10] of both static and hydrodynamic quantities in a melt. In the case of concentrated solutions, screening will not be operative at lengthscales below the mesh size  $\xi$ , but will hold at larger scales. But these contain the lengthscales of entanglement effects, so in the solution case also, coarse-grained local dynamics are expected to follow the Rouse model locally. Even in concentrated polymeric fluids the dynamics of chains (and subchains) below a characteristic molecular weight are not strongly coupled to the presence of other chains. The central assumptions are:

- **Gaussian Chains:** in which the force on a subchain segment  $n$  is the net entropic force from its neighbors. In the continuum representation we have adopted, this is equivalent to a thermodynamic force at each point of the chain of  $-\frac{\partial}{\partial n} \left( \kappa \frac{\partial \mathbf{R}}{\partial n} \right) = -\kappa \frac{\partial^2 \mathbf{R}}{\partial n^2}$  with  $\kappa = \frac{3k_B T}{b^2}$ .

- Local drag: the drag force on a subchain segment comes from frictional drag against background without long-range hydrodynamic effects of backflow (this works in melts where all long-range mediated backflows are screened). This force is  $\zeta \frac{\partial \mathbf{R}}{\partial t}$  with  $\zeta$  a drag coefficient per segment.
- Brownian motion: a random force  $\mathbf{f}(n, t)$  acts on each subchain with correlation times much faster than any polymer dynamics to be modeled by the theory.

The monomeric drag constant  $\zeta_0$  will parametrize all our theoretical models, setting the timescale for both Rouse and, subsequently, entangled, motion. The balance of entropic, drag and random forces on the chain of  $N$  subchains is the Rouse equation:

$$\zeta_0 \frac{\partial \mathbf{R}}{\partial t} = \frac{3k_B T}{b^2} \frac{\partial^2 \mathbf{R}}{\partial n^2} + \mathbf{f}(n, t). \quad (16)$$

The noise on each subchain is related to its frictional drag by the generalized Einstein relation as above:

$$\langle \mathbf{f}(n, t) \mathbf{f}(m, t') \rangle = 2\zeta_0 k_B T \mathbf{I} \delta(n - m) \delta(t - t'). \quad (17)$$

The Rouse dynamical Eq. (16) is diagonalized by the transformation:

$$\mathbf{R}(n, t) = \mathbf{X}_0(t) + 2 \sum_{p=1}^{\infty} \mathbf{X}_p(t) \cos\left(\frac{p\pi n}{N}\right). \quad (18)$$

The  $\mathbf{X}_p(t)$  are the time-dependent amplitudes of the “Rouse modes” of the polymer chain. These are just the (vector amplitude) Fourier modes of the chain path  $\mathbf{R}(n, t)$  with respect to the arclength coordinate  $n$ . The key result for us is the time correlation function of the mode amplitudes, which is:

$$\langle \mathbf{X}_p(t) \mathbf{X}_q(t') \rangle = \mathbf{I} \frac{k_B T}{k_p} \delta_{pq} e^{-|t-t'|/\tau_p} \quad \text{with } k_p = \frac{6k_B T p^2 \pi^2}{Nb^2}. \quad (19)$$

Each mode has its own relaxation time  $\tau_p = \zeta/k_p$  that decrease rapidly (as  $1/p^2$ ) with mode index  $p$ . The longest of these relaxation times  $\tau_1 = \frac{\zeta N^2 b^2}{3\pi^2 k_B T}$  has special significance. It is known as the *Rouse Time*, and often given the notation  $\tau_R$ . It is the time for relaxation of the overall shape of the molecule (it is the relaxation time of the amplitude of the normal mode with fewest nodal points,  $\cos \frac{\pi n}{N}$ ), and is also the time for a Gaussian Rouse chain to diffuse its own radius of gyration.



**Polymer Physics, Figure 4**

A Rouse chain changes its configuration (from solid to dashed curve) locally but not globally in times shorter than  $\tau_R$

What does the local motion of this model chain look like? We expect for short intervals that the chain contour may have adjusted locally, but retain a very similar global configuration (see Fig. 4). In order to answer this question, and to compare with local diffusion probes of NSE and NMR on unentangled dynamics, we need to calculate the correlation function  $\phi_n(t) \equiv \langle (\mathbf{R}(n, t) - \mathbf{R}(n, 0))^2 \rangle$  that describes the mean displacement over time of monomers on the chain. Summing over the independent normal modes gives

$$\phi_n(t) = 6D_{CM}t + \frac{Nb^2}{3\pi^2} \left( \frac{t}{\tau_R} \right)^{1/2} \alpha.$$

Here  $\alpha = \frac{1}{2} \int_0^\infty z^{-3/2} (1 - e^{-z}) dz \simeq 1.77$  is a purely numerical constant. The result is remarkable: each monomer executes an “anomalous” or sub-Fickian diffusion, such that its mean square displacement goes as  $t^{1/2}$  rather than  $t$  (as for ordinary diffusion). This behavior persists until times longer than the Rouse time, after which each monomer is carried by the (faster) center of mass motion of the whole molecule. This anomalous diffusion is simply a consequence of chain connectivity: the further a monomer travels under Brownian motion, the greater is the length of chain that must be correlated with it and the greater the effective drag over that lengthscale.

The (deviatoric) stress formula

$$\sigma_{ij} = \frac{3k_B T}{b^2} \mathbb{C} \left\langle \frac{\partial R_i}{\partial n} \frac{\partial R_j}{\partial n} \right\rangle$$

we derived above Eq. (13) leads, via representation in terms of the Rouse modes [97], to an expression for the time-dependent modulus function following a step strain  $G(t) = \sigma_{xy}(t)/\gamma$ . Each mode decays back to equilibrium

anisotropy with its own characteristic time. The power-law distribution of modes then gives:

$$G(t) = \frac{\mathbb{C}k_B T}{N} \sum_p e^{-2p^2 t/\tau_R} \approx \mathbb{C}k_B T \left( \frac{t}{\tau_1} \right)^{-1/2} e^{-t/\tau_R}. \quad (20)$$

So we find that, until a final crossover to an exponential decay beyond the Rouse time, the Rouse model has a relaxation modulus which is a power-law of  $G(t) \sim t^{-1/2}$ . Note that the longest relaxation time scales with molecular weight as  $N^2$ , but the viscosity scales as  $\mathbb{C}k_B TN$ . This is because at the longest relaxation time that sets the value of the viscosity, the stress is carried only by the lowest Rouse mode; the density of these modes is just one per chain, or  $\frac{\mathbb{C}}{N}$ .

Beautiful generalizations of such dynamical scaling behavior arise in the case of marginally overlapped fractal clusters, such as those arising from gelation transitions. Long-range hydrodynamic interactions are effectively screened, but entanglement effects are negligible. So now the Rouse Eq. (16) generalizes to solving the eigenvalues of the Laplacian (right hand side of (16)) on a fractal cluster with high degree of branching. Locally (for high frequency modes) this is indistinguishable from the case of linear chains, but at longer wavelengths than the chain length between branch points, the effective dimensionality of the cluster rises. Just as in many cases it is possible to represent the mass distribution with an effective dimensional exponent  $D = 1/\nu$ , so the Laplacian eigenfunctions of a self-similar branched object possess a “spectral dimension”  $d_s$  and the density of states in  $k$ -space can be written

$$g(k)dk \cong k^{d_s-1}dk. \quad (21)$$

The combination of the scaling dimension of individual clusters, and the self-similarity of the molecular weight distribution via the Fisher exponent yields a generalized power-law for stress relaxation [35]. For Rouse dynamics (local friction) and hyperscaling the result depends only on the fractal dimension of the clusters

$$G(t) \simeq \frac{kT}{b^3} \left( \frac{t}{\tau_0} \right)^{-\frac{3}{D+2}}. \quad (22)$$

The value of the dynamic exponent  $z$  in  $G(t) \sim t^{-z}$  for three-dimensional percolation is therefore predicted to be  $z \cong 0.66$ . Experiments on critical gels of unentangled chains confirm this [41].

## Entangled Polymer Dynamics

Richer behavior still emerges in the realm of inter-chain *topological* effects that dominate in fluids where chain overlap is very strong. The polymer melts of industrial polymer processing are very highly overlapped on the molecular level, where it becomes immediately apparent that molecular relaxation processes controlling elastic stress are prolonged to very long times indeed. The classic “relaxation modulus”  $G(t)$  measuring stress linear-response to a step strain records a “plateau” value before a terminal relaxation time that increases rapidly with molecular weight, in strong contrast to the power-law decays discussed above. Experiments restricted to the timescales of the plateau are hardly able to distinguish between the polymer melt and a rubber, in which the chains are permanently cross-linked to each other at very rare points, sufficiently for each chain to be permanently immobilized from large-scale diffusion. Conceptually, the absent “cross-links” were replaced in the minds of engineers and physicists alike by “entanglements” [13]. These loosely-defined objects were assumed to represent the topological constraint that covalently-bonded molecular chains may not pass through each other. The effective distance between these objects could be calculated, employing rubber elasticity theory,

$$G_N^{(0)} = k_G \frac{RT\rho}{M_e}$$

(with the constant  $k_G = 1$  for “affine” and  $1/2$  for “junction fluctuation” models of elasticity) to deduce the degree of polymerization between entanglements  $N_e$ , or the equivalent “entanglement molecular weight”,  $M_e$ . The number  $N_e$  consistently turned out to be of order  $10^2$ , indicating a length-scale for an “entanglement spacing” of 50–100 Å, depending on the particular chemistry. This is highly significant for us, because it shows that small chains on the threshold of feeling topological interactions are real polymers, already long enough to show to a good approximation all the universal properties of statistical connected chains. It also suggests that the role of topology in highly-entangled ( $N \gg N_e$ ) polymer fluids has the potential to be treated universally. Further evidence of universality in entanglements came from experiments in which the polymers were diluted to a volume fraction  $\phi_p$  by a compatible solvent. The apparent entanglement molecular weight  $M_e \sim \phi_p^{-\alpha}$  where the scaling exponent  $\alpha \simeq 1$  [11].

Other experiments had pointed to the existence of a topological feature at this coarse-grained scale of structure. Careful measurements on rubbers of controlled synthesis had shown that the shear modulus was higher for



a network of long chains than a model incorporating cross-links alone would predict [14]. Other “trapped entanglements” on the same scale as the melt value of  $N_e$  seemed to contribute to the elasticity. Advanced theories of rubber elasticity have been able to treat rubber networks in terms of the two distinct constraints of physical cross-links and trapped entanglements [16,42]. A remarkable universality also emerged in measurements of the scaling of melt viscosity  $\eta$  on the molecular weight  $M$  of very many different polymer chemistries [13]:

$$\begin{aligned} \eta &\sim M^1 & M < M_c \\ \eta &\sim M^{3.4} & M > M_c \end{aligned} \quad (23)$$

For each material, a critical molecular weight,  $M_c$  emerged, above which the viscosity rises very steeply with molecular weight. Furthermore, within experimental error, this explicitly dynamical observation was linked phenomenologically to the essentially static measurements of the plateau modulus by the correlation

$$M_c \simeq 2M_e. \quad (24)$$

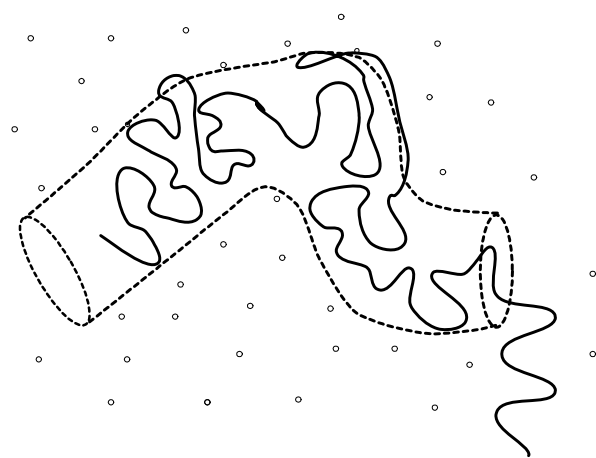
This connection between essentially dynamic ( $M_c$ ) and static ( $M_e$ ) experiments, observed over a wide range of chemistries, is strong evidence that topological interactions dominate both the molecular dynamics and the viscoelasticity at the 10 nm scale in polymer melts (and at correspondingly larger scales for concentrated solutions).

Without going beyond rheological measurements on bulk samples, there has long been other very strong evidence that molecular topology is the dominant physics in melt dynamics. This emerges from the phenomenology of “long chain branched” (LCB) melts. These materials, commonly used in industry, possess identical molecular structure to their linear cousins on the local scale, but contain rare molecular branches (they differ from the “polymeric fractals” discussed above in that their linear sections are long enough to be entangled). The density of branching varies from one branched carbon in every 10 000 to 1 in 1000. This level is chemically all but undetectable, yet the melt rheology is changed out of all recognition if the molecular weight is high enough [43]. Providing that  $M \gg M_e$ , the limiting low-shear viscosity may be much higher for the same molecular weight. Moreover in strong extensional flows the melt responds with a much higher apparent viscosity than in linear response. This phenomenon, vital for the stable processing properties of branched melts, is called “extension hardening”. The effect is all the more remarkable because in shear flows, branched, as well as linear, melts exhibit a lower stress

than would be predicted by a continuation of their linear response [44] (they are “shear-thinning”). A fascinating example of the difference between linear and branched entangled melts is well-known from flow-visualization experiments. The velocity field in a strong “contraction flow” of a linear polymer melt resembles that of a Newtonian fluid, while that of a branched polymer sets up large vortices situated in the corners of the flow field. Slight changes to the topology of the molecules themselves give rise to qualitatively different features in the macroscopic fluid response.

The most successful accounts of these phenomena have been given by the *tube model*. The idea is to deploy the theoretical physicists favorite strategy of replacing a difficult many-body problem with a tractable single-body problem in an effective field. In this case the “single body” is the single polymer chain, and the effective field becomes a tubelike region of constraint along the contour of the chain. The tube is invoked to represent the sum of all topological non-crossing constraints active with neighboring chains, and the tube radius,  $a$ , is of the order of the end-to-end length of a chain of molecular weight  $M_e$ . In this way, only chains of higher molecular weight than  $M_e$  are strongly affected by the topological constraints (see Fig. 5).

The tube was first invoked by Edwards [45] in an early model for the trapped entanglements in a rubber network. The consequences of the idea for dynamics were first explored by de Gennes [46], again in the context of networks. A free chain in a network would be trapped by neighbor-



**Polymer Physics, Figure 5**

A tubelike region of constraint arises around any selected polymer chain in a melt due to the topological constraints of other chains (small circles) in its neighborhood (diagram courtesy of R. Blackwell)

ing chains into tube of radius  $a$  defined by its own contour, suppressing motion perpendicular to the tube's local axis beyond a distance of  $a$ , but permitting both local curvilinear chain motions and center-of-mass diffusion along the tube. de Gennes coined the term “reptation” for this snake-like wriggling of the chain under Brownian motion. The theory gives immediately a characteristic timescale for disengagement from the tube by curvilinear center-of-mass diffusion. This disengagement time  $\tau_d$  is naturally proportional to the cube of the molecular weight of the trapped chain (this arises from combining the Fickian law of diffusive displacement of length  $L$  with time  $\tau$ ,  $\tau \sim L^2$ , recognizing that path length  $L \sim M$ , with one extra power arising from the proportionality of the total drag on molecular weight). Very significantly, de Gennes also realized that a tubelike confining field would endow a dangling arm, fixed to the network at one end, or belonging to a star-shaped polymer in a network, with exponentially slow relaxations. In this topology reptation would be suppressed by the immobile branch point [47], and only exponentially-rare retractions of the dangling arm would disengage it from its original tube (see Fig. 6 below). In the late 1970s, S.F. Edwards and M. Doi developed the tube concept into a theory of entangled melt dynamics and rheology for monodisperse, linear chains [97], finding extensions to flow instabilities [48], blends [49] and polymers of controlled architecture that go beyond the star topology [50] to H-polymers [51] and combs [52].

Fully atomistic simulations of polymer melts are now able to examine entanglement effects. It is now possible to conduct molecular dynamics simulations of, for example, elastically-connected Lennard–Jones polymers that

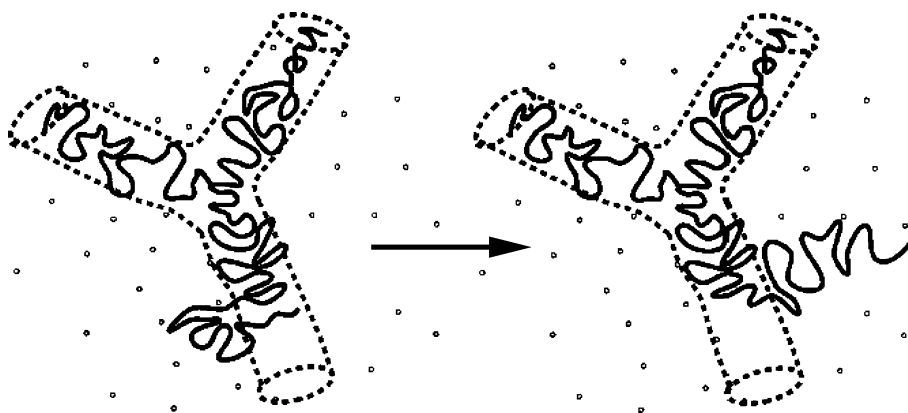
contain 50 chains each of 10 000 monomers well into the regime in which entanglements dominate the dynamics [53]. This technology is now at the point at which direct comparisons to experimental results such as NSE is now possible. The other advantage of large simulations is that they may mimic the “ideal” experiment in which everything may in principle be measured. This has been exploited in tests of fundamental theories of entanglements (see below) [54].

The growing quantity of data on branched molecules of controlled molecular weight and topology has provided severe tests of the tube concept at a level beyond that probed by linear chains [40]. The hierarchical nature of configurational relaxation at the molecular level in particular has been turned from speculation into orthodoxy. In the simplest case of entangled star polymers, the theory suggests that chains escape from their confining tubes not by reptation, which is suppressed by virtue of the immobile branch point, but by a process of *arm retraction*, present but largely eclipsed in the case of linear polymers (see Fig. 6).

The effect on the viscosity of replacing linear molecules with those of identical molecular weight, but of star topology, is striking: now

$$\eta \sim e^{\alpha(M_a/M_e)} \quad M_a > M_c \quad (25)$$

is the dominant form of the molecular weight dependence (where  $M_a$  is the molecular weight of the dangling *arm*), rather than  $\eta \sim M^{3.4}$  (in the case of linear polymers the entire effect of these fluctuations is to change the apparent exponent of this relation from 3 to 3.4 up to a high molecular weight of order  $10^3$  entanglements, where it sat-



Polymer Physics, Figure 6

The process of arm retraction predicted by the tube model for the case of dangling entangled arms, as from the branch point of a star polymer. Unlike in reptation, reconfiguration of the outer parts of the arm occurs many times for one relaxation of deeper segments

urates at the “bare reptation” value of 3). In entangled melts with repeated levels of branching, the retraction dynamics of outer levels generate the slower retraction of inner branches in a hierarchical way. The most advanced application of this process has been on materials that combine the complexity of the gelation-point critical ensemble (see Subject. “[Single Chain Dynamics](#)” above) with the physics of entanglement. The broad rheological spectrum of branching chemistry that builds molecules by grafting back repeatedly previously synthesized molecules [55] is captured by these calculations. When the percolation transition is approached by cross-linking entangled polymers, there exists a measurable region in which mean-field statistics ( $D = 4$ ) actually hold. Stress relaxation from such an ensemble is logarithmic in this regime, but is closely modeled by a dynamical scaling form in which the dynamic exponent  $z$  is small (being inversely proportional to the number of entanglements between branch points) [56].

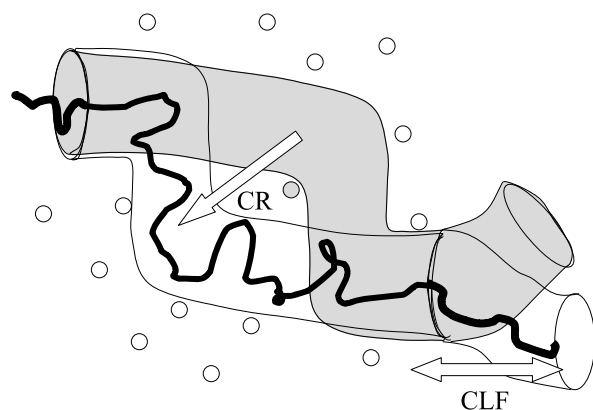
The wealth of experimental and simulation data has sharpened the theoretical picture. Without exploding with new parameters, it has been possible to capture, in a single model, modes of entangled motion beyond pure reptation. In linear response *contour length fluctuation* (CLF), the Brownian fluctuation of the length of the entangle-

ment path through the melt, modifies early-time relaxation. Similarly the process of *constraint release* (CR), by which the reptation of surrounding chains endows the tube constraints on a probe chain with finite lifetimes, contributes to the conformational relaxation of chains at longer times. Both the processes of CLF and CR contribute to the quantitative understanding of linear rheology, such that the  $\eta \sim M^{3.4}$  law is no longer a mystery [57,58], but much of the newer data still need to be examined quantitatively as sensitive tests of the detailed physics, and many puzzles remain. These two additional processes are visualized in Fig. 7.

### Non-linear Flow of Polymeric Fluids

In strong deformations the additional processes of *chain stretch*, *chain retraction*, and *branch-point withdrawal* emerge on the level of single chains (the latter exclusively in the branched case), and *convective constraint-release* (CCR) at the level of co-operative motion [59,60,61]. The most advanced formulation of the tube model for linear polymers keeps the coarse grained coordinates of the chain, and allowing CCR events to generate local Rouse jumps of the tube [61]. The idea is to retain full information about average chain trajectories instead of working indirectly with dynamic equations for the stress and orientation tensor. This approach also allows quantitative predictions about the single chain scattering function  $S(q)$ , and to develop a *local* description of CCR events. The main assumptions of the first version of the theory (valid when there is no chain stretch) are: (i) that CCR operates locally in reorienting chain segments both into and away from the flow direction, and (ii) that neither the number of entanglements per chain  $Z = M/M_e$  nor the tube diameter  $a$  changes. The first assumption endows the tube itself with a Rouse-like motion in which the local hopping rate is coupled to the global deformation rate via a single new parameter. The second (constant length) assumption introduces a difference from ordinary Rouse-chain motion, and limits the range of validity at first (but see below) to  $0 < \dot{\gamma} < 1/(\tau_e Z^2)$ .

No single set of variables will be able to diagonalize the essential entangled modes of motion, namely (i) chain reptation, (ii) chain retraction, (iii) tube-length fluctuation and the new mode (iv) Rouse-tube motion. However, the theory is conventionally cast in a real-space notation for the tube trajectory  $\mathbf{R}(s, t)$  and its tangent curve  $\mathbf{R}' \equiv \frac{\partial \mathbf{R}}{\partial s}$ , functions of curvilinear distance  $s$  from along the tube and time  $t$ . Our chains are monodisperse containing  $Z$  entanglements of tube diameter  $a$ . The (stochastic) equation of



**Polymer Physics, Figure 7**

A cartoon of the processes of contour length fluctuation (CLF) and constraint release (CR) on a linear polymer in a constraining tube. In CLF the chain end retracts via longitudinal fluctuations of the entangled chain, but without requiring center-of-mass (reptation) motion. Re-extension of the chain end may explore new topological constraints, reconfiguring the tube. In CR, an entanglement with a neighboring chains (shown hatched) may disappear, allowing effective conformational relaxation of that part of the tube, again without reptation of the test chain itself. In both cases the former tube configuration is shown dark, the new, light

motion becomes [62]

$$\begin{aligned} \mathbf{R}(s, t + \Delta t) = & \kappa \mathbf{R} \Delta t + \mathbf{R}(s + \Delta \xi(t), t) \\ & + \Delta t \left( \frac{3\nu}{2} \frac{\partial^2 \mathbf{R}}{\partial s^2} + \mathbf{g}(s, t) \right) \\ & + \Delta t \lambda \left( \frac{Z}{2} - s \right) \frac{\partial \mathbf{R}}{\partial s} + 3D_c Z \frac{(\mathbf{R}'' \mathbf{R}')}{|\mathbf{R}'|^2} \mathbf{R}'. \end{aligned} \quad (26)$$

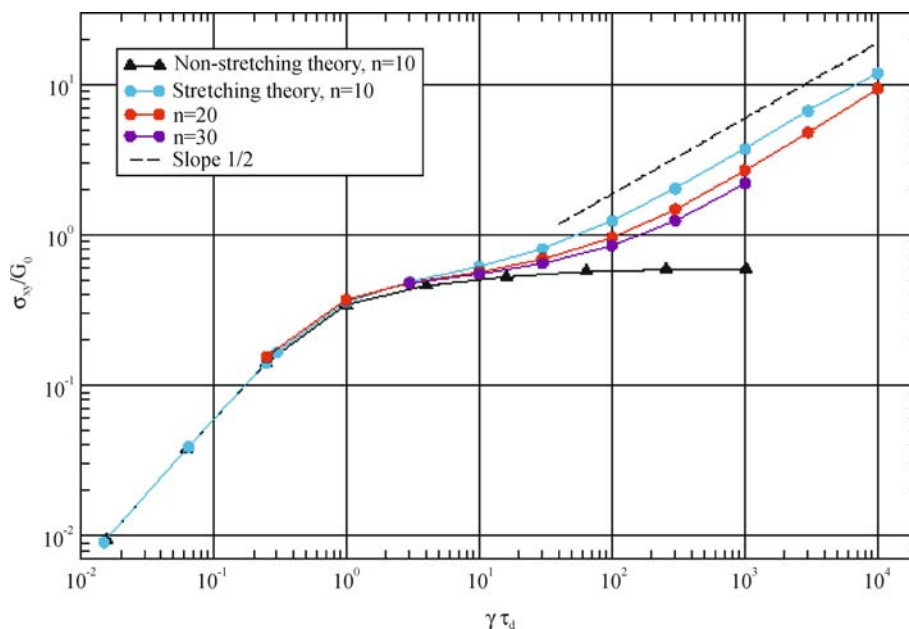
The terms of this formulation describe, in order, affine convection of the tube, reptation, CR Rouse motion of the tube, retraction of the chain within the tube and chain stretch. This model predicts in shear flow a near-plateau of  $\sigma_{xy}(\dot{\gamma})$  between  $\dot{\gamma}\tau_d$  and  $\dot{\gamma}\tau_R$ , and an increase proportional to  $\dot{\gamma}^{1/2}$  beyond that (so that the “shear-dependent viscosity”  $\eta(\dot{\gamma}) \sim \dot{\gamma}^{-1/2}$  (see Fig. 8). Both this feature, and a prediction of strong extension hardening at rates faster than the inverse chain stretch time, have been quantitatively matched to experiment.

The strongest tests of the predictions for chain conformations have been performed in “neutron flow mapping” experiments [64,65]. Here a partially-deuterated (for neutron scattering contrast) polymer melt of monodisperse or controlled architecture, is passed continuously through a complex flow field such as a contraction, bounded by windows transparent to both laser illumination and ther-

mal neutrons. The stress field is measured in optical birefringence, while single chain conformations are reported by the neutron small angle scattering. Scanning the apparatus across the neutron beam allows the experiment to probe regions of the flow with varying strain histories. The experiments are compared to calculations in which the model of (26) is used both to calculate the flow field itself, then the scattering function of chains subjected to the stream lines that flow through the neutron beams.

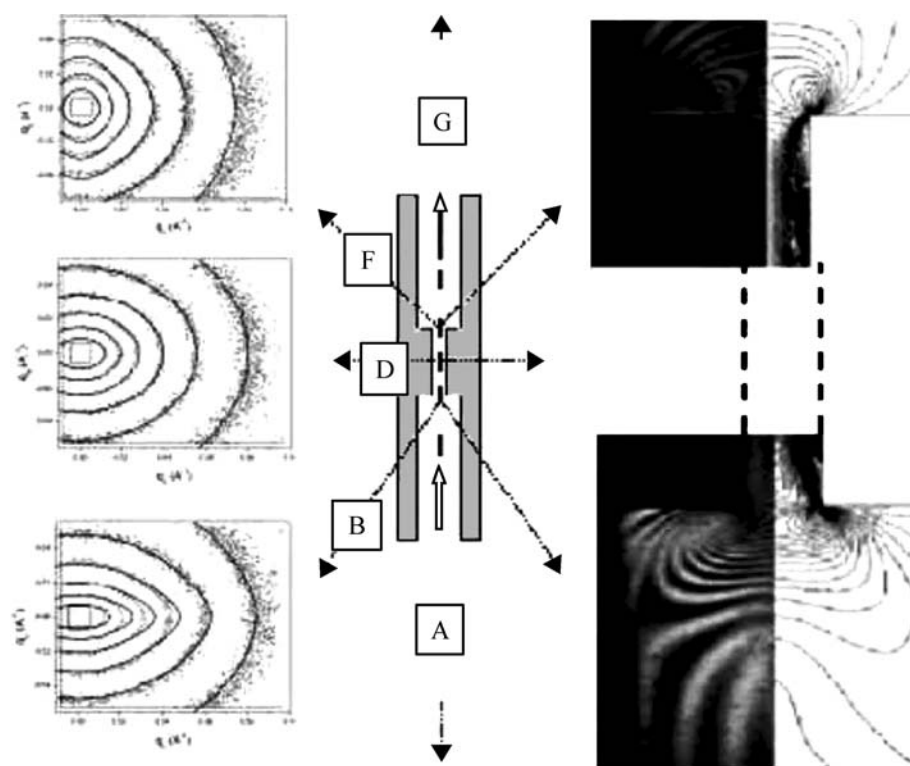
Figure 9 shows the results of one such experiment. The key result is that relaxation to equilibrium structure takes place in the flow at timescales that depend on the chain lengthscale examined. It also shows that no model containing just one viscoelastic relaxation time is even qualitatively able to account for the non-linear physics of entangled melts. The simplest must possess at least two relaxation times, corresponding to chain stretch (fast – by Rouse motion) and chain orientation (slow – by reptation). At the purely phenomenological level of the stress tensor the emergent property is a rapidly-relaxing trace, and a slower traceless part to the stress.

At a more approximate level, it is possible to combine the complexities of non-linear response and complex branched topologies in entangled melts. All the linear and non-linear molecular processes of chains in tube-like entanglement fields are present, together with one additional process, that of “branch point withdrawal”. When a strand



**Polymer Physics, Figure 8**

Predictions of the local CCR model with chain stretch using  $c_v = 0.1$  for values of  $Z = 10, 20, 30$ . A comparison to the non-stretching version is given

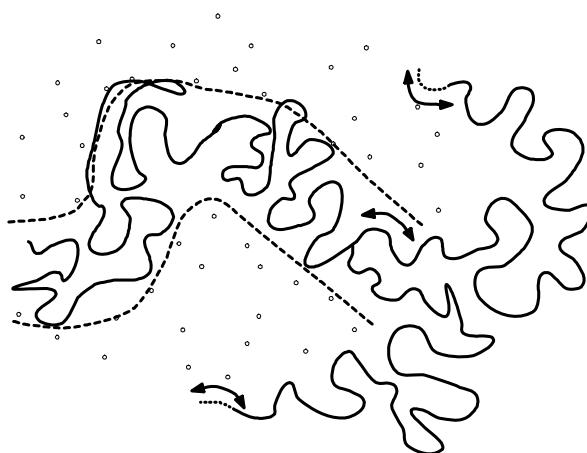


**Polymer Physics, Figure 9**

A monodisperse linear entangled melt in a neutron flow mapping experiment. Flow is from bottom to top. Single chain scattering functions from windows B, D and F are compared with contour predictions on the *left*. Chain configurations relax during passage through the channel from an initial strain. On the *right*, contours of principal stress (*left*) are compared with calculations (*right*)

connecting two branch points is very highly stretched, the finite size of the clusters it connects (in contrast to the network case) becomes apparent, and one of the clusters may topologically collapse into the deforming tube of the central strand. In the simplest case of the H-polymer the two outer arms are drawn into the tube of the “cross-bar” segment (see Fig. 10).

There are consequences of the process for both chain conformation (scattering) and stress response (rheology). If the retracting arms are labeled in a scattering experiment, very strong signal enhancement is seen in the direction of the deformation [51,66]. At the same time, the growing extensional stress, until that point resembling that of a cross-linked network, reaches a near-plateau. The effect of this at the level of the process engineering is to endow strongly branched melts with both extension hardening (leading to flow stabilization) and good processability. A useful modeling tool for branched melts has been derived from the ideal “pom-pom” architecture [67]. This generalization of the H-polymer allows the number of arms attached at either end of the cross-bar,  $q$ , to vary. So  $q$  becomes a molecular parameter controlling



**Polymer Physics, Figure 10**

The process of branch point withdrawal: a segment with greater than equilibrium tension pulls attached dangling arms some distance into its own tube, thus shortening their effective entangled path length



the degree of strain-hardening. Creating multiple modes from this model allows accurate models to be built for commercial, highly-branched melts [68,69,70]. Computing in complex geometries with these models has successfully predicted features of the flows particular to branched melts, such as recirculating vortices upstream of a contraction and very persistent sheets of high birefringence downstream from re-entrant corners, sometimes known as “stress-fangs” [71].

### Multi-Phase Polymeric Fluids

In the foregoing, the spatial structure of polymeric fluids has remained essentially homogeneous, all heterogeneities remaining at the level of entanglements, or correlation volumes, in solution. The hidden reason for this is that we have considered systems containing just one chemistry of monomer. Yet one of the great attractions of polymers is their tendency to develop complex spatial structures when different chemistries are combined, either in the case of entire chains, when we create *polymer blends*, or within single chains, referred to as *co-polymers*. In the latter case, especially interesting structures arise when the distinct monomers are correlated in sequence along the chain, forming *block co-polymers*. The simplest example would be a string of  $n$  monomers of chemistry  $A$  followed by  $m$  of chemistry  $B$ . This is an  $A$ - $B$  diblock co-polymer. The essential reason for the sensitivity to local chemistry is that the entropy of spatial translation per monomer in a polymeric fluid is extremely low. The usual Van't-Hoff term  $S_{\text{trans}} = -kT \log \phi$  is divided by the degree of polymerization of the chain so that it effectively competes with the mild Van der Waals dominated enthalpic interaction between monomer units, that tend to favor proximity of identical chemical units: polymers tend therefore towards demixing. In blends the demixing competes with the slow, viscoelastic dynamics we discussed in the last section, creating spatially complex morphologies that are determined kinetically. In the case of block co-polymers, the demixing occurs locally, confined to spatial scales of the block subchain radii. The demixing now competes with the chain elasticity we discussed in Sect. “Single Polymer Chain Physics”. The connection of polymer sequence with emergent structure illustrates the high potential information content of a macromolecule. It is an example of the genotype-phenotype pattern developed to a much higher degree in the case of the DNA-embedded genetic code. This field, like that of controlled architecture dynamics, is an area of polymer physics where the complex emergent phenomena have required a parallel implementation of careful synthetic chemistry [72], experimental physics [73]

and advanced theoretical techniques [75] to explore, and is growing extremely rapidly.

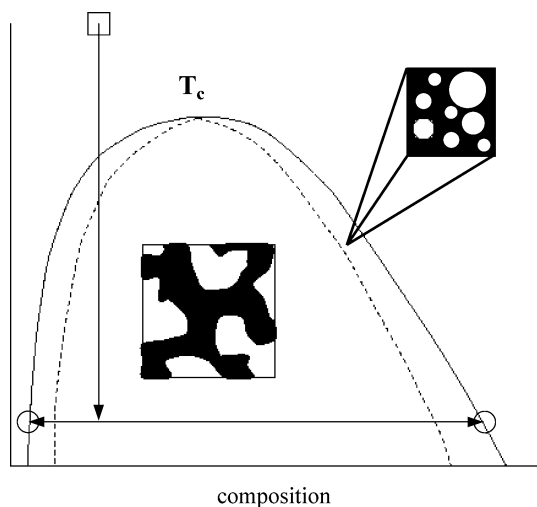
### Polymer Blends

The starting point for a conceptual understanding of the physics of polymer blends is a mean-field model for the free energy of mixing of two species such that one occupies a total volume fraction  $\phi$ . Known as the Flory–Huggins free-energy [17], it balances the entropy of mixing of the two components against the energy difference of mixed and demixed states:

$$\Delta F_{\text{mix}} = k_B T \left[ \frac{\phi}{N_A} \ln \phi + \frac{1-\phi}{N_B} \ln (1-\phi) + \chi \phi (1-\phi) \right]. \quad (27)$$

The control parameters for this theory are the two molecular weights and the (temperature dependent) interaction (“Flory”) parameter  $\chi$ . The whole system is therefore three-dimensional: the monomer interaction is best normalized with the mean molecular weight so that  $\chi \bar{N}$  is the essential parameter that controls interaction strength. Then  $N_A/N_B$  becomes the asymmetry parameter while  $\phi$  controls the composition. The possibility of phase separation means that  $\Delta F_{\text{mix}}$  can be minimized by forming regions with different values of  $\phi$  if the curvature  $\partial^2 \Delta F_{\text{mix}} / \partial \phi^2$  is anywhere less than zero. This in turn occurs for all  $\chi > \chi_c$ . Phase separation then occurs for all systems whose mean composition falls between the two minima of  $\Delta F_{\text{mix}}(\phi)$ , a region that broadens as  $\chi$  moves further (as temperature is changed) from  $\chi_c$ .

This behavior is mapped in Fig. 11. The region between the two curves (binodal and spinodal) of the plot corresponds to compositions and temperatures where the curvature of  $\Delta F_{\text{mix}}(\phi)$  is positive, producing a fluid that is *locally* stable to composition fluctuations though globally unstable to phase separation. In this case droplets of the separated phase have to nucleate, giving an initially disconnected morphology. Within the spinodal curve on the other hand, the fluid is unstable to local and infinitesimal changes in composition, so that natural thermal fluctuations of composition are amplified. The presence of a fastest-growing wavelength leads to a connected (or “spinodal”) morphology for most of this region. Since the final minimum in free-energy is total phase separation, whatever intermediate structure evolves eventually coarsens with time with well-known growth laws depending on whether hydrodynamics or diffusion is dominating [76]. This beguilingly simple map hides a great deal of latent complexity, however. More recently, the recognition that the spinodal process leads to a morphology dependent

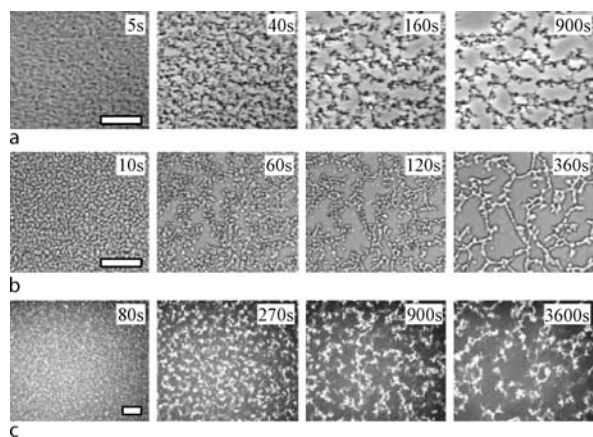


**Polymer Physics, Figure 11**

Regions of phase-separation in polymer blends with the Flory–Huggins model. Phase separation occurs within the binodal curve (solid curve), but is unstable locally only within the spinodal (dashed curve)

on both the local free energy and the distribution of initial fluctuations has generated experimental and theoretical explorations of the wider control space that includes trajectories across the composition diagram [77], variations on boundary conditions and viscoelastic effects [78]. A striking example is given by the “target” structures that result from kinetic trajectories that spend some time in the nucleation region of the composition space before (cooling) into the spinodal region. During the first period, small nuclei are allowed to form but sustain limited growth. Instead, the growth occurs in the locally unstable region using the nuclei as sources of unstable fluctuations. Because the Fourier wavelet decomposition of a spherical nucleus with a sharp boundary consists of spherical waves centered on the nucleus, it is these that are amplified by the growth process, or reverse-diffusion, and structures of nested spheres of varying composition appear during it. Since phase separating polymers typically also have a glass transition temperature, it is possible to quench the composition structure at any point in the phase separation.

Viscoelasticity may play a modifying role at both early and late stages of the phase separation. At early times, any dominant viscoelastic mode, such as reptation, can undergo mode-mixing with the dominant phase separation time to produce to two-timescale process of separation [79]. This may result in non-monotonic concentration growth with time, since of the two resulting modes one typically decays while the other grows. At later stages, a strong viscoelasticity in at least one phase causes the het-



**Polymer Physics, Figure 12**

Confocal imaging of viscoelastic phase separation from [78] showing the time development of structure from a dilute polymer solution, b protein solution, c concentrated polymer solution

erogeneous fluid to retain the mutual connectivity of both phases for much longer than in Newtonian fluids [78].

Figure 12 illustrates the high degree of universality of this effect, contrasting dilute and concentrated polymer solutions with a protein solution in which the protein molecules behave more as colloidal particles than polymer chains. Although the structures coarsen, they retain connectivity of the minority phase rather than suffer it breaking into droplets. A similar effect is generated by the natural composition-dependence of mobility. It is unlikely that the mobilities of the two demixing species will remain independent of the local composition, since the natural drag constants within the two final demixed phases will typically differ. The phase of lower mobility then becomes kinetically quenched earlier than that of higher mobility. This nonlinearity affects the connectivity of the morphology at long times [80].

### Block Co-Polymers

The demixing tendencies of polymers of different chemistries becomes very rich when it competes with the feature at the heart of polymer physics itself: that of connectivity. By connecting a chain of *A*-monomers to one of *B*-monomers they are forced by the spatial correlations that chain connectivity induces into proximity. Within a mean field picture, the repulsive interaction in the free energy density  $k_B T \chi \phi_A \phi_B$  (from the last term in (27)) competes with the entropic chain elasticity  $F_{\text{elas}} = k_B T (R_A^2/N_A b^2 + R_B^2/N_B b^2)$ . In contrast to previous sections, we consider first the case of dense chains

(melts), then solutions and finally the effects within single chains of such controlled chemical architecture.

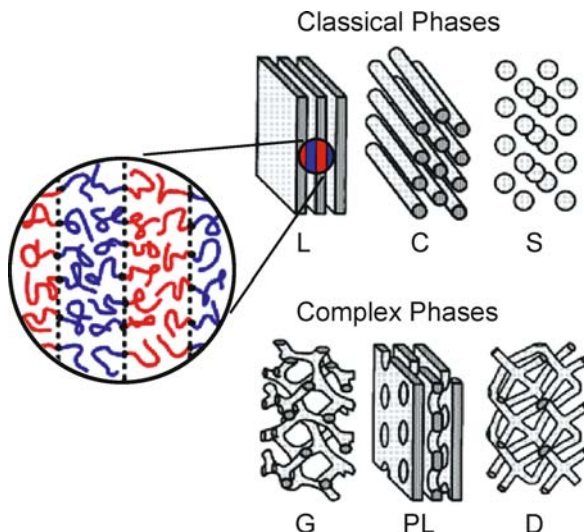
Consider first the case of the *diblock* co-polymer just outlined. In a melt, when the interaction parameter  $\chi$  is large enough, the chains must locally segregate into regions rich in one or other of the polymers. The molecules themselves clearly minimize their contribution by situating their midpoints where the two chemistries join at the interface between the two regions. Topology dictates that at most two *A*-chains may span an *A*-rich region, so consequently the length scale of the morphology must be of the same order as the radius of gyration of the chains. In the symmetric case the system spontaneously forms *lamellae* of alternating composition whose width increases as the interaction parameter does though with a weak power law of  $\chi^{1/6}$ . The width of the interface region in which both *A* and *B* monomers are present decreases with  $\chi$ . In a fully quantitative “strong segregation theory” of this physics [81] the lamellum width  $L$  and interface width  $w$  obey

$$L = 2 \left( \frac{8}{3\pi^4} \right)^{1/6} N^{2/3} b \chi^{1/6}; \quad w = \frac{2b}{(6\chi)^{1/2}}. \quad (28)$$

But additional physics comes into play as soon as the two blocks are of different length, for now the entropy of confinement of the chains in the plane of the interface does not balance (we did not consider that in the above), with the result that the interface tends to curve away from the domains containing the longer chains. Other more complex morphologies become candidates for the minimum in free energy: periodic cylinders (C) and spheres (S) as well as the more exotic forms of the gyroid (G), perforated lamellae (PL) and double-diamond lattices (D) of Fig. 13.

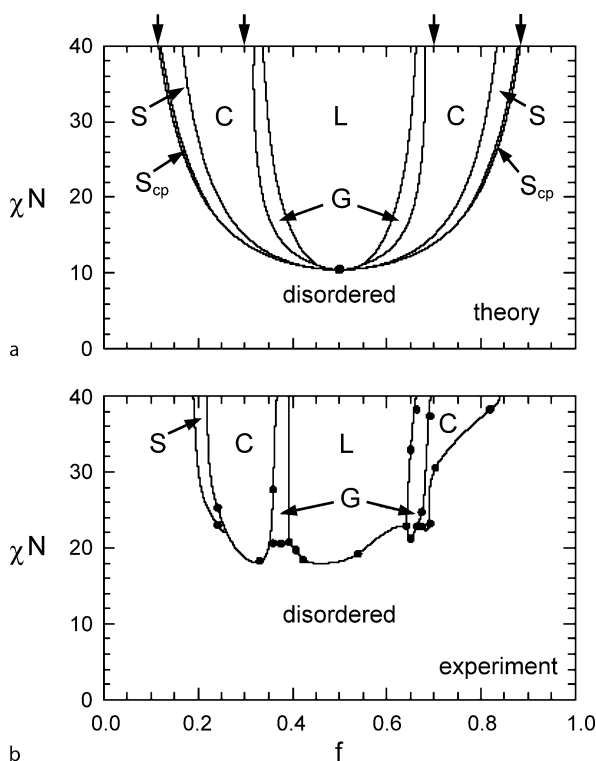
Experiments on carefully synthesized block co-polymers, especially the model polystyrene-polyisoprene system, have mapped out the morphology diagram in terms of interaction (via temperature, as in blends) and composition of the diblock. Calculations were first performed near the critical point, where segregation of the two species is only weak, and the free energy can be expanded in powers of the difference of the local mean concentrations  $\phi = (\phi_A - \phi_B)$  [83], but can be extended into strong segregation by a fully continuous self-consistent mean field theory [82] that ignores only the fluctuations in composition. Results and comparison with experiment are shown in Fig. 14.

Clearly the calculations are qualitatively, and in some aspects quantitatively in agreement, especially far from the disordered (fully mixed) state. However, fluctuations clearly have an important effect near the boundary. The



Polymer Physics, Figure 13

Potential candidate morphologies for block co-polymer melts, from [82]



Polymer Physics, Figure 14

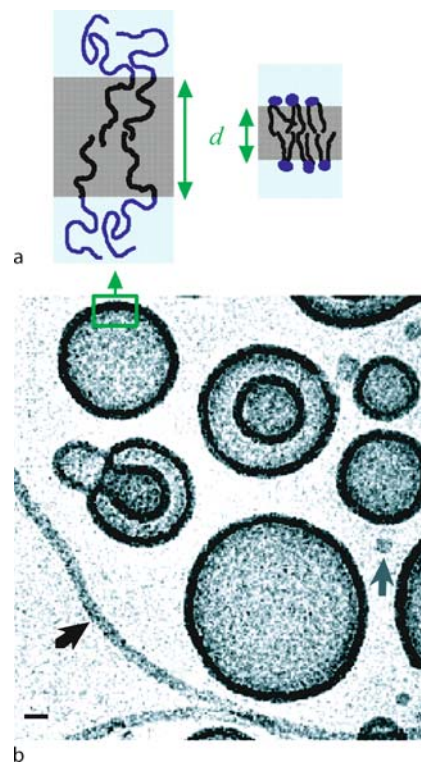
The morphology map for diblocks as a function of interaction parameter and composition by a self-consistent field calculation and **b** experiment on PS-PI diblocks. From [82]

topology of the diagram at the critical point is actually changed by them, stabilizing a finite region of the lamellar phase. Quite recently, field-theoretical methods for computing the full effect of fluctuations in the partition function of block co-polymers have been introduced [84]. Other natural extensions include increasing the complexity of the chemical structure of the block co-polymer. Taking only one step to the tri-block permits a very wide class of morphologies that mix those of the diblock, so that for example spheres of *A* may decorate cylinders of *B* within a matrix of *C*. Such materials are by no means only of academic interest, as they constitute the fundamental technology behind toughened plastics, in which the minority, rubbery phase prevents crack propagation within a majority glassy phase [85]. We can begin to see the sequence of the chain corresponding to an information content (“genotype”) that implicitly codes for the morphology (“phenotype”) of the emergent morphology.

Block copolymers in solution, either of a low molecular weight species or of a simple homopolymer, may act as polymeric versions of the familiar surfactants. Their self-assembly picks out the same structures as we have already seen in the context of bulk microphase separation (lamellae, cylinders and spheres) with now the difference that these self-assembled structures exist in isolation rather than in a regular array. The corresponding structures are vesicles, wormlike micelles and spherical micelles [86].

Examples are given in Fig. 15. These phases have a visibly complex form, since they are largely determined kinetically: because of the long timescales for diffusion and collision of the individual micelles, and of the high activation energy for mutual rearrangement, merger and breaking, true equilibrium is very hard to reach in these systems. So vesicles of widely different radius may coexist, together with nested structures and complexes. The transitions between linear and spherical structures, which may be driven by changes in temperature or pH depending on the chemistry of the polymers, may exhibit a rich kinetics in which cylindrical structures emerge from spherical and vice versa.

We finally consider the generalization of the coil-collapse transition we saw in the case of dilute chains in poor solvents in Sect. “Single Polymer Chain Physics”. When all monomers are identical in their interaction with the solvent the form of the resultant globule will be on average spherical, together with natural thermal fluctuations of the interface. When the polymer contains blocks of heterogeneous interaction with solvent, as well as self-interaction, the globule may become a much more complex object. Experiments on diblocks indicate that a range of



**Polymer Physics, Figure 15**

**Schematic of chain configuration in the wall of a block co-polymer vesicle and its architectural control (a) and transmission electron micrograph of polymeric vesicles (b), from [86]**

non-spherical geometries may be generated by controlling the architecture of the primary chain [87]. Calculations balancing chain stretch, mutual interaction and surface tension against the solvent [88] indicate that it is theoretically possible, using only two monomers, to code for transitions between near-spherical to prolate and finger-like forms of the collapsed globule. More complex features such as budded and pearled structures also emerge from the same mean-field level of theory that successfully treats block co-polymer melts. This single-molecule form of the coding of emergent phenotype from the information coded as a polymer sequence is very suggestive of nature’s own method of constructing the functional single molecules of enzymes, motors and cellular structures. For proteins are “just” co-polymers (of a possible 20 amino acid monomer set) that code in their sequence for absolutely specific forms of their collapsed state. The protein-folding problem has generated a huge literature [89], although rather little of it actually exploits polymer physics, with its natural high order of dimensionality and degrees of freedom [90] to understand this ultimate refinement of the art of coded morphology.



## Future Directions

Several current themes suggest a rich future for polymer physics, notwithstanding the experience of history that the richest veins of research to come will be those currently unforeseen. The general area of biomimetic, or perhaps better, bio-suggestive polymer physics is bound to be an area of growth. Biology has already mastered the art of synthesizing single polymers that act as self-assembling machines, chemical reactors and separation systems. Artificial polymers that act in these ways are certainly possible, and may not need to be as exactly structured as proteins are in order to deliver function. We have already seen that theoretical schemes exist for designing block co-polymers that will collapse into pre-determined shapes and forms [88]. Future designs of active versions of protein-like polymers may also go beyond the chemical “fuel” of ATP dephosphorylation, perhaps using light as both an energy and signaling source. Optically-activated mechanical transitions in polymers have already been demonstrated [91]. Effectively equally fast response can be elicited from pressure changes. In the long term one might hope for advanced therapeutics from this route, especially in combination with polymer-based encapsulation systems such as triggered micelles of block co-polymers.

Structural materials properties also have many things to learn from evolution. As it becomes possible to moderate and control microstructure, both in terms of crystallinity and microseparation, so polymeric nano-composites will be able to realize combinations of strength and stiffness currently existing as ideals [74].

A related direction brings polymer physics into biological research directly. Dynamic neutron scattering by neutron spin echo [92], as well as advanced NMR relaxation techniques will assist the current move towards exploring the role of dynamics in molecular biological function on a similar footing to the achievements in the area of structure. Even thermal dynamics is beginning to be recognized as a generator of functions such as signaling, in an analogous way to the emergence of rubber elasticity from the same source [93]. Technologically, the use of artificial polymers to create scaffolds for tissue engineering will require a balance of local biochemical interaction and global mechanical and topological structure formation.

The key underpinning science of biocompatible and biomimetic polymers is the control of self-assembly. Already there are a number of “suprachemical” options of non-covalent, reversible polymerization [94]. Playing with nature’s alphabet of peptide-forming amino acids, that naturally self-assemble via main-chain hydrogen bonds, gives a rich system in which chirality becomes a new con-

trol parameter for the equilibrium structure [95]. Combining main-chain self-assembly with side-chain functionality may prove to open up new classes of functional polymeric materials.

The growth of conducting and semi-conducting polymeric materials within a burgeoning new sector of the electronics industry is already well under way. But this area has been driven largely by technology, and much of fundamental science remains to be understood, especially at the level of many chain, materials physics [96]. The changing demands of the world’s energy economy are bound to put pressure on developments of organic, polymer-based photovoltaic materials, as well as lightweight polymer gel energy storage. The combination of information-processing and advanced structural properties within new polymeric materials is a tempting prospect.

## Bibliography

### Primary Literature

1. Kuhn W (1936) *Koll Z* 76:258
2. Guth E, Mark H (1934) *Monatshfte Chemie* 65:93
3. Flory PJ (1953) *Principles of Polymer Chemistry*. Cornell University Press, Ithaca
4. Zimm BH, Stockmayer WH (1949) *J Chem Phys* 17:1301
5. Edwards SF (1976) The configuration and dynamics of polymer chains. In: Balian R, Weill G (eds) *Molecular Fluids*. Gordon and Breach, London, pp 151–208
6. Zimm BH (1956) *J Chem Phys* 24:269
7. Rouse PE (1953) *J Chem Phys* 21:1272
8. Zinn-Justin J (1993) *Field Theory and Critical Phenomena*. Clarendon Press, Oxford
9. De Gennes PG (1986) *Scaling Concepts in Polymer Physics*. Cornell University Press, Ithaca
10. Edwards SF (1966) *Proc Phys Soc Lond* 88:265
11. Adam M, Delsanti M (1984) *J Phys France* 45:1513
12. Verdier PH, Stockmayer WH (1962) *J Chem Phys* 36:227
13. Ferry JD (1986) *Viscoelastic Properties of Polymers*. Wiley, New York
14. Treloar LRG (1975) *The Physics of Rubber Elasticity*. Clarendon, Oxford
15. James H, Guth E (1947) *J Chem Phys* 15:669
16. Ball RC, Doi M, Edwards SF, Warner M (1981) *Polymer* 1010:22
17. Koningsvelt R, Stockmayer WH, Nies E (2001) *Polymer Phase Diagrams*. Oxford University Press, Oxford
18. Feynman R (1965) *Hibbs Quantum Mechanics and Path Integrals*. McGraw Hill, Kogakusha
19. Perkins TT, Smith DE, Larson RG, Chu S (1995) Stretching of a single tethered polymer in a uniform flow. *Science* 268:83
20. Marko JF, Siggia ED (1995) Stretching DNA. *Macromolecules* 28:8759
21. Edwards SF (1965) *Proc Phys Soc* 85:613
22. Mazur J, McCrackin FL (1968) *J Chem Phys* 49:648
23. de Gennes PG (1972) *Phys Lett* 38A:339–341
24. des Cloiseaux J (1981) *J Phys* 42:635
25. des Cloiseaux J, Jannink G (1990) *Polymers in Solution*. Oxford University Press, Oxford



26. Rubinstein M, Colby RH (2003) *Polymer Physics*. Oxford University Press, Oxford
27. Dobrynin AV, Rubinstein M (2005) Theory of polyelectrolytes in solutions and at surfaces. *Prog Polym Sci* 30(11):1049–1118
28. Manning GS (1969) Limiting Laws and Counterion Condensation in Polyelectrolyte Solutions I. Colligative Properties. *J Chem Phys* 51(3):924–933
29. Ha BY, Liu Andrea J (1998) Effect of Non-Pairwise-Additive Interactions on Bundles of Rodlike Polyelectrolytes. *Phys Rev Lett* 81:1011
30. Daoud M et al (1975) *Macromolecules* 8:804
31. Noda I et al (1981) *Macromolecules* 14:668
32. des Cloiseaux J (1975) *J Phys* 36:281
33. Deam RT, Edwards SF (1976) *Phil Trans Roy Soc A* 280:317–353
34. Castillo HE, Goldbart P (2000) *Phys Rev E* 62:8159
35. Cates ME (1985) Excluded volume and hyperscaling in polymeric systems. *J Phys Lett* 46:L837–L843
36. Jian T, Vlassopoulos D, Fytas G, Pakula T, Brown W (1996) *Colloid Polym Sci* 274:1033
37. Higgins JS, Benoit H (1994) *Polymers and Neutron Scattering*. Clarendon Press, Oxford
38. Klein PG, Adams CH, Brereton MG, Ries ME, Nicholson TM, Hutchings LR, Richards RW (1998) *Macromolecules* 31:8871
39. Watanabe H (1999) *Prog Polym Sci* 24:1253
40. McLeish TCB (2002) Tube Theory of Entangled Polymer Dynamics *Adv Phys* 51:1379–1527
41. Lusignan CP et al (1995) *Phys Rev E* 52:6271
42. Rubinstein M, Panyukov S (2002) *Macromolecules* 35:6670
43. Small PA (1975) *Adv Polym Sci* 18:1
44. Meissner J (1975) *Pure Appl Chem* 42:551
45. Edwards SF (1967) *Proc Roy Soc London* 92:9
46. de Gennes PG (1971) *J Chem Phys* 55:572
47. de Gennes PG (1975) *J Phys (Paris)* 36:1199
48. McLeish TCB, Ball RC (1986) *J Polym Sci Polym Phys Edn* 24:1755; McLeish TCB (1987) *J Polym Sci Polym Edn Phys* 25:2253
49. Watanabe H, Kotaka T (1984) *Macromolecules* 17:2316
50. Pearson DS, Halfand E (1984) *Macromolecules* 17:888
51. McLeish TCB, Allgaier J, Bick DK, Bishko G, Biswas P, Blackwell R, Blotti  re B, Clarke N, Gibbs B, Groves DJ, Hakiki A, Heenan R, Johnson JM, Kant R, Read DJ, Young RN (1999) *Macromolecules* 32:6734–6758
52. Inkson NJ, Graham RS, McLeish TCB, Groves DJ, Fernyhough CM (2006) *Macromolecules* 39:4217–4227
53. P  tz M, Kremer K, Grest GS (2000) *Europhys Lett* 49:735
54. Everaers R (1998) *Eur Phys J B* 4:341
55. Das C, Inkson NJ, Read DJ, Kelmanson Mark A, McLeish TCB (2006) *J Rheol* 50:207
56. Das C, Read DJ, Kelmanson MA, McLeish TCB (2006) *Phys Rev E* 74:011404
57. Milner ST, McLeish TCB (1998) *Phys Rev Lett* 81:725
58. Likhtman AE, McLeish TCB (2002) *Macromolecules* 35:6332–6343
59. Marrucci G, Non-Newt J (1996) *Fluid Mech* 62:279
60. Mead DW, Doi M, Larson RG (1998) *Macromolecules* 31:7895
61. Likhtman AE, McLeish TCB, Milner ST (2000) *Phys Rev Lett* 85:4550
62. Graham RS, Likhtman AE, Milner ST, McLeish TCB (2003) *J Rheol* 47:1171
63. Fetters LJ, Lohse DJ, Graessley WW (1999) *J Polym Sci Polym Phys Edn* 37:1023
64. Bent J et al (2003) *Science* 301:1691–1695
65. Graham RS (2006) *Macromolecules* 39:2700–2709
66. Heinrich M, Pyckhout-Hintzen W, Allgaier J, Richter D, Straube E, McLeish TCB, Wiedenmann A, Blackwell RJ, Read DJ (2004) Small-Angle Neutron Scattering Study of the Relaxation of a Melt of Polybutadiene H-Polymers Following a Large Step Strain. *Macromolecules* 37:5054–5064
67. McLeish TCB, Larson RG (1998) *J Rheol* 42:81
68. Inkson NJ, McLeish TCB, Groves DJ, Harlen OG (1999) *J Rheol* 43:873
69. Blackwell RJ, McLeish TCB, Harlen OG (2000) *J Rheol* 44:121
70. Graham RS, McLeish TCB, Harlen OG (2001) *J Rheol* 45:275
71. Lee K, Mackley MR, McLeish TCB, Nicholson TM, Harlen OG (2001) Experimental observation and numerical simulation of transient stress fangs within flowing molten polyethylene. *J Rheol* 45:1261–1277
72. Ruokolainen J, Mezzenga R, Fredrickson GH, Kramer EJ, Hustad PD, Coates GW (2005) Morphology and thermodynamic behaviour of syndiotactic polypropylene-poly(ethylene-co-propylene) block polymers prepared by living olefin polymerization. *Macromolecules* 38(3):851–60
73. Adhikari R, Michler GH (2004) Influence of molecular architecture on Morphology and Micromechanical behaviour of styrene/butadiene block copolymer systems. *Prog Polym Sci* 29:949–86
74. Leibler L, Ajdari A, Mourran A, Coulon G, Chatenay D (1994) Ordering in macromolecular systems. Springer, Berlin
75. Matsen MW (2005) In: Gompper G, Schick M (eds) *Soft Condensed Matter*. Wiley-VCH, Berlin
76. Onuki A (2002) *Phase Transition Dynamics*. Cambridge University Press, Cambridge
77. Henderson IC, Clarke N (2004) *Macromolecules* 37:1952–1959
78. Tanaka H, Araki T, Koyama T, Nishikawa Y (2005) Universality of viscoelastic phase separation in soft matter. *J Phys Cond Mat* 17:S3195–S3204
79. Clarke N, McLeish TCB, Pavawongsak S, Higgins JS (1997) Viscoelastic Effects on the Phase-Separation of Polymer Blends. *Macromolecules* 30(15):4459–4463
80. Mao Y, McLeish TCB, Teixeira PIC, Read DJ (2001) Asymmetric landscapes of early spinodal decomposition. *Eur Phys J E* 6:69–77
81. Semenov AN (1985) *Sov Phys-JETP* 61:733
82. Matsen MW (2002) The standard Gaussian model for block copolymer melts. *J Phys Cond Mat* 14:R21–R47
83. Leibler L (1980) *Macromolecules* 13:1602
84. Fredrickson GH (2007) Computational field theory of polymers: opportunities and challenges. *Soft Matter* 3:1329–1334
85. Leibler L (2005) Nanostructured plastics: Joys of self-assembling. *Prog Polym Sci* 30:898–914
86. Discher DE, Eisenberg A (2002) *Science* 297:967–973
87. Govorun EN, Ivanov VA, Khokhlov AR, Khalatur PG, Borovinsky AL, Grosberg AYU (2001) *Phys Rev E* 64:R40903
88. Khokhlov AR, Semenov AN, Subbotin AV (2005) Shape transformations of protein-like copolymer globules. *Eur Phys J E* 17:283–306
89. Dinner AR,   ali A, Smith LJ, Dobson CM, Karplus M (2000) Understanding protein folding via free-energy surfaces from theory and experiment. *TIBS* 25:331–339
90. McLeish TCB (2005) Protein Folding in High-Dimensional Spaces: Hypergutters and the Role of Nonnative Interactions. *Biophys J* 88:172–183

91. Hugel T, Holland NB, Cattani A, Moroder L, Seitz M, Gaub HE (2002) Single-Molecule Optomechanical Cycle. *Science* 296:1103–1106
92. Bu Z, Biehl R, Monkenbusch M, Richter D, Callaway DJE (2005) Coupled protein domain motion in Taq polymerase revealed by neutron spin-echo spectroscopy. *PNAS* 102:17646–17651
93. Hawkins RJ, McLeish TCB (2004) Coarse-Grained Model of Entropic Allostery. *Phys Rev Lett* 93:098104
94. Lehn JM, Mascal M, DeCian A, Fischer J (1990) *J Chem Soc Chem Commun* pp 479
95. Davies RPW, Aggeli A, Beevers AJ, Boden N, Carrick LM, Fishwick CWG, McLeish TCB, Nyrkova I, Semonov AN (2006) Self-assembling  $\beta$ -Sheet Tape Forming Peptides. *Supramol Chem* 18:435–443
96. Barford W (2005) *Electronic and Optical Properties of Conjugated Polymers*. Oxford University Press, Oxford
97. Doi M, Edwards SF (1986) *The Theory of Polymer Dynamics*. Oxford University Press, Oxford

## Books and Reviews

- Larson RG (1999) *The Structure and Dynamics of Complex Fluids*. Clarendon Press, Oxford
- Fredrickson GH (2006) *The Equilibrium Theory of Inhomogeneous Polymers*. Oxford University Press, Oxford

## Polymers, Non-linearity in

KOH-HEI NITTA

Division of Material Sciences, Graduate School  
of Natural Science and Technology, Kanazawa University,  
Kanazawa, Japan

## Article Outline

Glossary  
Definition of the Subject  
Introduction  
Topological Analysis of Branched Molecules  
Ideal Chain Models  
Chain Statistics of Nonlinear Polymers  
Chain Dynamics of Nonlinear Polymers  
Future Directions  
Bibliography

## Glossary

**Connectivity** Connectivity extends the concept of adjacency and is essentially a form (and measure) of concatenated adjacency. If it is possible to establish a path from any vertex to any other vertex of a graph, the graph is said to be connected; otherwise, the graph is disconnected. A graph is totally disconnected if there

is no path connecting any pair of vertices. This is just another name to describe an empty graph or independent set.

**Distance** The distance  $d_{uv}$  between two vertices  $u$  and  $v$  in a graph is the length of a shortest path between them. When  $u$  and  $v$  are identical, their distance is 0. When  $u$  and  $v$  are unreachable from each other, their distance is defined to be infinity  $\infty$ .

**Ideal chain** The interactions between adjacent structural units along the chain are called short-range interaction, whereas the interactions between units which are far removed from each other along the chain are called long-range interaction. Note that the long-range interactions are typically short-range in space. An ideal chain is the unperturbed chain without the intramolecular long-range interaction between structural units where only short range interaction is considered. A basic theoretical model for ideal flexible chains is the Gaussian chain, which assumes a number of ideal beads with intramolecular distance between them following a Gaussian distribution.

**Long chain branching** The clearest definition of this is that to be long a branch needs to have a molecular weight at least greater than the entanglement molecular weight which is defined in terms of the plateau modulus, which can be directly measured from the linear viscoelasticity of the polymer.

**Nonlinear polymer** These are branched and cross-linked polymers which contains some polyfunctional units. This term is reserved for functionalities exceeding two and the branches formed from the polyfunctional units sufficiently are long or large.

**Polymer** A chain like molecule make up of repetition of a particular atomic group joined together by covalent bonds. The chain-like molecule is usually called the polymer if the entanglement and intertwining interaction occurs in the melt state and in the concentrated solution. The basic unit of this sequence is called the ‘structural unit’, and the number of units in the sequence is the degree of polymerization.

**Tree** A tree is a connected acyclic simple graph. A vertex of degree 1 is called a leaf, or pendant vertex. An edge incident to a leaf is a leaf edge, or pendant edge. (Some people define a leaf edge as a leaf and then define a leaf vertex on top of it. These two sets of definitions are often used interchangeably.) A non-leaf vertex is an internal vertex. Sometimes, one vertex of the tree is distinguished, and called the root. A rooted tree is a tree with a root. Rooted trees are often treated as directed acyclic graphs with the edges pointing away from the root.

## Definition of the Subject

Branching formation is known from the beginning of polymer chemistry and modern synthesis methods make it possible to prepare a great variety of nonlinear polymers with specific branching structure. Branched polymers are nowadays becoming more and more important so that analytical techniques to reveal the role of branching in macroscopic properties of polymers are desired. This article demonstrates that the graph-theoretical approaches are most effective when providing topological and physical insights into the nonlinearity in polymer architecture. Thus, the problems of the dynamics and statistics of any branched molecule are shown to be completely reduced to the problem of the eigen-polynomial of graph, resulting in that various ideas and concepts, thus, obtained from the graph theory can be applied directly to the topological analysis for architecture in nonlinear polymers. This will lead to a new kind of paradigm in polymer science and engineering.

## Introduction

A polymer is a chain-like molecule that comprises huge number of repeating structural units or atoms connected by chemical bonds. Modern polymerization techniques for the preparation of chain-like molecules can produce chain branching and these branched polymers possess practically important properties and specific phenomena involving processability and rheological properties [19,34,50,55,87,109] which cannot be reached for linear polymers.

Branching formation was suspected almost from the beginning of the study of polymer chemistry in the 1930s. Flory [41] first pointed out the possibility of the occurrence of branching in the free-radical polymerization of diene monomers. The occurrence of branching had been well recognized by the early 1940s and since then, many theoretical and experimental studies related to the effects of branching on the properties of polymers have been carried out. In 1953, several important papers [2,7,108,115] showed that polyethylene materials produced by the free-radical process have not only a significant number of short branches but also long-chain branches.

In the 1960s, several research groups found that the “living polymer” procedure of anionic polymerization makes it possible to synthesize definitely known branched polymers [88]. During the 1970s and 1980s in addition to random branching that occurs in industrial polymerization, more rational synthetic methods based on living anionic polymers have facilitated the preparation of a great variety of polymers with specific branching structures such as star-shaped polymers having several branches attached

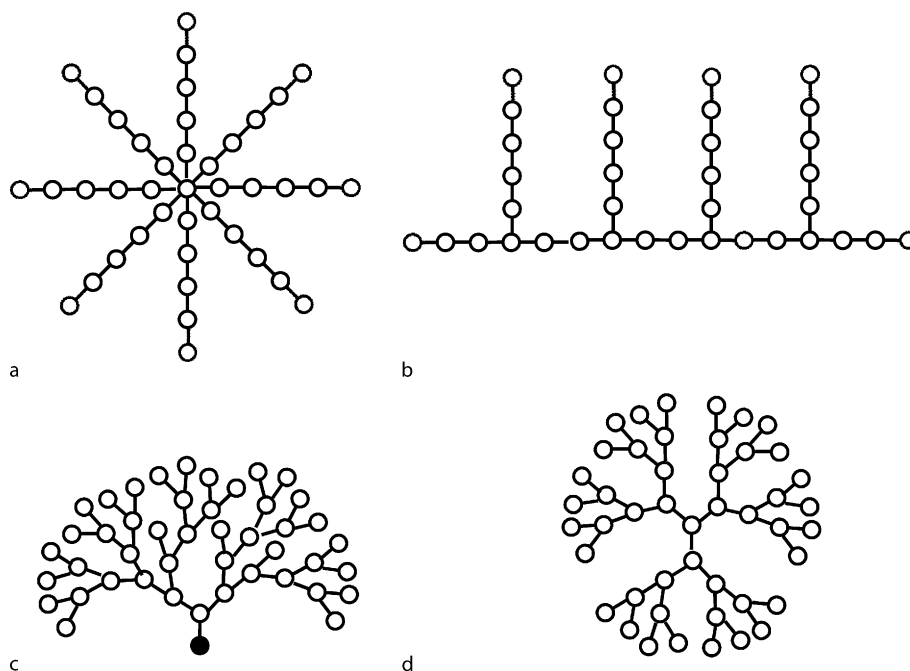
to a single polyfunctional branching point [58] and comb-shaped polymers having a given number of branching points randomly distributed along a backbone [93,124]. Comb-shaped polymers with branching points of functionality greater than three are also sometimes called polymeric brushes [80]. During the period from the end of the 1980s to the 1990s, it has been possible to build structures possessing regular “hyper-branched” polymers [74] or starburst polymers with radial symmetry, which is generally called a “dendrimer” [127]. The physical and chemical properties of such dendritic polymers, which essentially differ from those of not only linear polymers but also comb- and star-shaped polymers, enables unpredictable wide applicability [49]. These specific branched polymers are illustrated in Fig. 1.

In order to clarify the effects of long chain branching on polymer properties, we need some quantitative measures that reflect molecular branching: one of the quantitative factors is the average size of isolated single polymers characterized through the measurement of dilute solution properties. The smaller expansion in the space of isolated branched molecules as compared with isolated linear molecules of the same molecular weight is the basis for the most fundamental methods for estimating branching as demonstrated by Zimm–Stockmayer [139].

For the purpose of estimating branching, therefore, considerable effort has been extended to obtain the chain dimensions such as the mean square radius of gyration  $\langle s^2 \rangle$  estimated by performing the light scattering (LS) measurements and/or the mean Stokes radius  $r_H$ , estimated by using the intrinsic viscosity  $[\eta]$  data. In addition, a gel permeation chromatograph (GPC) [20,26,60] is one of the most popular devices used for the fractionation of a polymer according to the volume dimension  $[\eta]M$  where  $M$  is the molecular weight. Therefore, the effort to relate  $[\eta]$  to  $\langle s^2 \rangle$  or  $r_H$ , has been continuing for a long time by a combination of GPC and LS or  $[\eta]$  techniques, although they cannot satisfactorily describe the structural details of branched polymers such as the number of branches, the branch length, and their position along the backbone.

Under ordinary conditions, in a dilute solution, there are significant long-range intramolecular correlations that “perturb” the conformation of the chains. However, under special conditions in the  $\Theta$  state [84], the effects of the excluded volume of a structural unit vanish and the dimensions of the macromolecule adopt their unperturbed values which are appropriate for evaluating the branching degree based on conformational statistics.

Consequently, extensive research has been carried out in calculating the various conformational properties corresponding to well-defined branching architectures. How-



Polymers, Non-linearity in, Figure 1

Structural images of a regular star; b regular comb; c hyperbranch; d dendrimer

ever some mathematical difficulties often arise in applying statistical treatments to unperturbed branched polymers. The Monte Carlo simulation methods [6,48] have been powerful computational tools in the treatment of branched chains more complicated than those generally treated by statistical methods. The simulations suggest that intrinsic properties of polymers are influenced not only by the number of branches but also by the branch length and their position along the backbone. Quantitative data concerning the local dynamics and local structure can be obtained from the molecular dynamics simulation studies. However, it is hard to provide the topological or physical insights into these computer simulation results.

The specification of a linear polymer requires only one parameter; i. e. the degree of polymerization or molecular weight; however, additional key parameters are necessary for the specification of any branched polymer. For understanding molecular branching from a general point of view, systematic studies on the effect of branching on the physical properties in the frame of a homologous series of polymeric materials are desirable. For this purpose, it is also necessary to develop the topological or graph-theoretical methods for analyzing the branching and/or skeletal nature of polymers such as the branch length, the number of branches, and their positions.

The final goal of our work is to set up a rigorous molecular theory for linear as well as nonlinear flexible polymers and to give a description and a reliable prediction of the materials properties in the melt and solid states. This article demonstrates that the graph-theoretical approaches are most effective when establishing such a universal framework for polymers with any structural architecture.

This article deals with the topological nature of tree-like chain molecules as typical nonlinear polymers. This article is organized as follows: The graph-theoretical approach for characterizing branched molecules is described in brief in Sect. “[Topological Analysis of Branched Molecules](#)”. In Sect. “[Ideal Chain Models](#)”, the chain conformational statistics in the unperturbed state is discussed in the framework of ideal chain models based on the spring-beads and rod-beads models. The graph-theoretical approach to chain statistics is presented in Sect. “[Chain Statistics of Nonlinear Polymers](#)”. The relations between the conformational statistics of ideal chains and the graphic representation are also presented. In Sect. “[Chain Dynamics of Nonlinear Polymers](#)”, it is shown that the dynamics of various types of branched chains are obtained by solving the polynomial equation derived from the graph-theoretical representations. The final section provides a brief summary of this article and future problems.

### Topological Analysis of Branched Molecules

An alkane molecule is one of the most basic organic compounds, and it is a set of carbon and hydrogen atoms that are connected to one another by covalent bonds. The topological analysis of a branched alkane begins with a drawing where the atoms are depicted as points and the bonds linking them are depicted as straight lines.

If only the carbon atoms of alkanes are depicted as points and their hydrogen atoms are omitted, there is a one-to-one correspondence between these isomers and the drawings whose points have at most four neighbors. This type of drawings is called the chemical graph [122]. In the graphs, the hydrogen atoms do not normally play a major role in discriminating the structure of a molecule and in enumerating the isomers of alkanes. It was major problem for chemists to predict the number of isomers of any alkane on the basis of simple graphical constructions. The enumeration of the chemical isomers, in particular the constructional isomers of alkanes, has been treated as the subject matter of a mathematical discipline known as graph theory since the pioneering work by Cayley [21] and Sylvester [123].

In graph theory, points (atoms) are generally referred to as *vertices* and lines (bonds) are referred to as *edges*, and the functionality of the atoms is called the “*vertex degree*”. We consider only the connected graphs in which every vertex has more than one neighbor, and no loops or multiple edges are involved. A hydrocarbon molecule can be represented as a tree graph  $G$ , in which a carbon atom (or vertex) and a bond (or edge) are arbitrarily num-

bered, and a digraph  $D$ , in which each edge of the graph  $G$  is arbitrarily directed, as exemplified in Figs. 2a and 2b. If the chemical graph is allowed to have a branch point with more than four degrees, the extended chemical graph, which may be called the molecular graph, could be used to describe the structural diversity in branched polymers by offering quantitative descriptors, called topological indices.

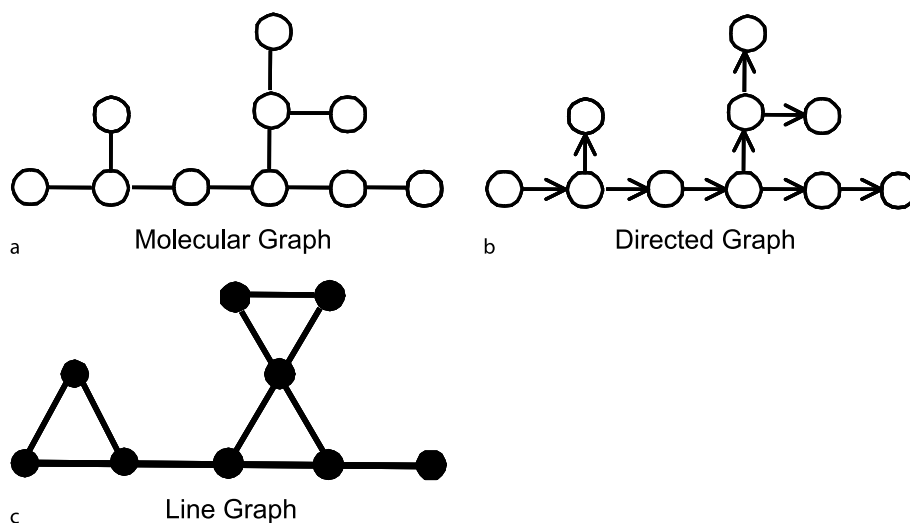
In graph theory, the algebraic expressions using several matrices reflecting the connectivity in graph  $G$  [69] are important devices for determining the topological feature of graphs, and the algebraic properties of the characteristic polynomials have been extensively examined. For graph  $G$ , the adjacency matrix  $A$  is the most fundamental matrix for the representation of graphs [69] and it is defined in graph theory as a square matrix with the following elements:

$$a_{ij} = \begin{cases} 1, & \text{if the vertexes } i \text{ and } j \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The matrix  $A$  is often called a topological matrix [68]. The order of  $A$  is identical with the total number of vertices in  $G$ . The adjacency matrix  $A$  of  $G$  has been useful for characterizing and encoding the skeletal structure of the corresponding molecules.

The characteristic polynomial of a molecular graph represents an important, even if not unique, molecular invariant. It is defined as

$$\Phi(A; \lambda) = \text{Det} |A - \lambda E|, \quad (2)$$



Polymers, Non-linearity in, Figure 2

Representation of a nonlinear molecule a the ordinary graph; b the digraph; c the line graph. The line graph (c) is transformed from the graph (a)



where  $\mathbf{E}$  is a unit matrix of the same order as  $\mathbf{A}$ . The list of eigenvalues of a matrix calculated from the characteristic polynomial is called the spectrum of the matrix which includes much quantitative information on topological nature of the molecules.

Another matrix describing a graph is the incidence matrix [70], which represents a linear mapping and determines the homology of the graph. The incidence matrix  $\mathbf{B} = (b_{ij})$  of a digraph  $D$  is defined by the following:

$$b_{ij} = \begin{cases} +1, & \text{if edge } j \text{ starts from vertex } i, \\ -1, & \text{if edge } j \text{ terminates in vertex } i, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The incidence matrix may be constructed for any graph having  $n$  vertices and  $m$  edges by setting up an  $n \times m$  matrix array; the rows and columns of the matrix then correspond to the vertices and the edges of the digraph, respectively. Also, the signless incidence matrix  $\mathbf{C} = (c_{ij})$  of an ordinary graph  $G$  can be defined and its elements are the absolute values of those of  $\mathbf{B}$ . Thus, the number 1 is inserted in the  $(i, j)$ th position in the array, if the  $j$ th edge is coincident with the  $i$ th vertex and all the other entries in the array are zeros.

Kirchhoff [75] discovered graphs while solving problems involving the calculation of currents in electrical networks where a connectivity matrix  $\mathbf{K} = \mathbf{C}^T \mathbf{C}$  was considered for admittance conductivity. The superscript “T” indicates the transpose of a matrix and a vector. He found the following formula relating the vertex adjacency matrix to the incidence matrix for tree graphs:

$$\mathbf{A}_L = \mathbf{K} - 2\mathbf{E}, \quad (4)$$

where  $\mathbf{A}_L$  represents the adjacency matrix for the line graph of  $G$ . The line graph  $L(G)$  of  $G$  is formed by replacing the edges of  $G$  by vertices in a manner such that the vertices in  $L(G)$  are connected whenever the corresponding edges in  $G$  are adjacent. The example of a line graph is shown in Fig. 2c. Furthermore, a combination of (4) and (2) gives

$$\Phi(\mathbf{K}; \lambda) = \Phi(\mathbf{A}_L; \lambda - 2). \quad (5)$$

Consequently, the eigenvalues  $\lambda_i$  of  $\mathbf{K}$  of a graph can be calculated from the eigenvalues  $\mu_i$  of  $\mathbf{A}_L$  of its line graph; i.e.  $\lambda_i = \mu_i + 2$ .

The topological analysis of graphs has widely been performed by the Laplacian matrices being related to both the adjacency and incidence matrices and they are defined by

$$\mathbf{L} = \mathbf{V} - \mathbf{A}, \quad (6)$$

$$\mathbf{L}^+ = \mathbf{V} + \mathbf{A}, \quad (7)$$

where  $\mathbf{V}$  is a diagonal matrix whose entries are the vertex degrees,  $\mathbf{L}^+$  is the signless Laplacian matrix [116,131], the entries of which is the absolute values of the entries of  $\mathbf{L}$ . The Laplacian matrices can be represented using the incidence matrices [24,130]

$$\mathbf{L} = \mathbf{B}\mathbf{B}^T, \quad (8)$$

$$\mathbf{L}^+ = \mathbf{C}\mathbf{C}^T. \quad (9)$$

The Laplacian matrix  $\mathbf{L}$  is in agreement with the Zimm matrix  $\mathbf{Z}$  which has been used for molecular dynamics of linear as well as branched polymers [137,138]. The details of the molecular dynamics are described in Sect. “Chain Dynamics of Nonlinear Polymers”. In addition, one can find the relation

$$\Phi(\mathbf{L}; \lambda) = \Phi(\mathbf{L}^+; \lambda), \quad (10)$$

indicating that the eigenvalues of the signless matrix  $\mathbf{L}^+$  are identical with those of the Laplacian matrix  $\mathbf{L}$ . Furthermore, comparing the characteristic polynomials of  $\mathbf{L}^+ = \mathbf{C}\mathbf{C}^T$  and  $\mathbf{K} = \mathbf{C}^T \mathbf{C}$ , we can find the following relation:

$$\Phi(\mathbf{L}^+; \lambda) = \lambda \Phi(\mathbf{K}; \lambda). \quad (11)$$

The eigenvalues of  $\mathbf{L}^+$  contain one zero eigenvalue and the non-zero eigenvalues of  $\mathbf{L}^+$  are identical with those of  $\mathbf{K}$ .

Another important invariant of a molecular graph is its distance matrix  $\mathbf{D}$  [70] which contains the shortest paths (distances  $d_{ij}$ ) between every pair of connected vertices. The entries in the distance matrix are related to the entries in powers of the adjacency matrix  $\mathbf{A}$ .

A secular determinant giving the Hückel molecular orbitals for the  $\pi$  electrons of an unsaturated hydrocarbon is reduced to the same form as the determinant of  $\mathbf{A}$  [3,56,57]. In other words, the problem of the Hückel orbital energies can be completely reduced to the eigenvalue problem of  $\mathbf{A}$ . Thus, the spectrum of the matrix  $\mathbf{A}$  yields the energy levels of molecular orbitals.

One of the earliest graph invariants or topological indices is known to mathematicians as the vertex number and to chemists as the carbon number. The carbon number is an appropriate index only for linear chain molecules; however, it is not well suited to branched molecules, which may have considerably different skeletal structures as compared to one another even if they have the same number of carbon atoms. Because there exists many different molecules with the same carbon number, the carbon number seems to be an index with low discriminating power.

Apparently, it is necessary to develop other indices that can effectively distinguish various types of isomers having different branching structures.

The first topological index capable of characterizing the branchedness of molecules was proposed by Wiener [133,134] to predict the boiling points of isomeric alkanes. The Wiener index  $W$ , named by Platt [102,103], is defined as the sum of the distances between any two carbon atoms in a hydrocarbon molecule:

$$W = \sum_{i=1}^{N-1} i k_i, \quad (12)$$

where  $k_i$  stands for the total number of pairs of atoms whose separation is  $i$ . Platt attempted to interpret this index that  $W^{1/3}$  is a sort of the mean molecular diameter [103].

The Wiener index of a hydrocarbon molecule is like the carbon number because it becomes generally larger for molecules with higher molecular weights; however,  $W$  also provides some measure of a molecule's branching structure. Since Wiener devised it, several other researchers found that the Wiener index correlated surprisingly well with properties for certain types of hydrocarbon molecules, as well as conjugated polymers [11,12,13,14,89,90] such as heat capacity, viscosity, surface tension, refractive index and electron energy.

Hosoya [71] found that  $W$  can also be obtained by the half sum of the off-diagonal elements of a distance matrix whose element  $d_{ij}$  is the number of edges for the shortest path between the  $i$ th and  $j$ th vertices: this not only offers an alternative method to determine  $W$  but also allows a particular extension of  $W$  to cyclic structures. The  $W$  is given by

$$W = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N d_{ij}, \quad (13)$$

where  $N$  is the total number of vertices. Rouvray examined the sum of the elements of the distance matrix, independently and considered it as a topological index [111]. The fact that Rouvray's index is equal to  $2W$  was soon recognized. Bonchev et al. [15] studied the Wiener index of any branched graph and succeeded in formulating structural features such as the branching point and the branch length.

The Laplacian matrix is a real symmetric matrix. The diagonalization of the Laplacian matrix for a graph  $G$  with  $N$  vertices produces  $N$  real eigenvalues,  $\lambda_1 > \lambda_2 > \dots > \lambda_N = 0$ , where the smallest eigenvalue is always zero. Let a graph  $G$  be a tree; then, the Wiener index of the tree can be obtained in terms of its Laplacian eigenvalues

as follows [91,95]:

$$W = N \sum_{i=1}^{N-1} \frac{1}{\lambda_i}. \quad (14)$$

In chemical graph theory, the distance matrix  $\mathbf{D}$  accounts for the bond interactions of atom in molecules. However, these interactions decrease as the distance between atoms increases so that the research groups of Balaban [73] and Trinajstić [104] respectively, proposed the reciprocal distance matrix whose entries are given by  $d_{ij}^{-1}$ . The reciprocal distance matrix enables the calculation of a Wiener index analog, as the half sum of its entries:

$$H = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^{-1}. \quad (15)$$

This index  $H$  is called the "Harary index". Here we generalize this index for its applicability to polymers as follows:

$$H_\varepsilon \equiv \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N d_{ij}^{-\varepsilon}. \quad (16)$$

Clearly, one can find  $W = H_{-1}$  and  $H = H_1$ .

Hosoya et al. [72] have introduced a topological index  $Z$ ; in the case of acyclic molecules,  $Z$  is equal to the sum of the absolute values of the polynomial coefficients of the adjacency matrix and it was used to describe the boiling points of various molecules. Recently, in a number of papers, the problem of molecular branching was related to the properties of the characteristic polynomial. Lovász and Pelikán [85] have found that the maximal eigenvalue of a tree graph is a fairly reliable measure of branching.

The Wiener index of a molecule is essentially based on the topological concept of *distance* so that it can be recognized as a measure of the molecular size rather than molecular shape and connectivity. Whereas Randić proposed a different type of index which is based on the topological concept of *degree*. The Randić index [107], or the connectivity index  $\chi$ , can be defined in terms of the atomic contributions and relates also to the elements of the normalized Laplace matrix. The vertex-connectivity index is defined as

$$\chi = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (d(v_i) d(v_j))^{-0.5}, \quad (17)$$

where  $d(v_i)$  is the degree of vertex  $i$ . Estrada [37] introduced the edge-connectivity index on the basis of the Randić index analog. These connectivity indices were found to be surprisingly correlated with the density, solu-

bility in water, molar volume, molar refractivity, and various types of biological responses.

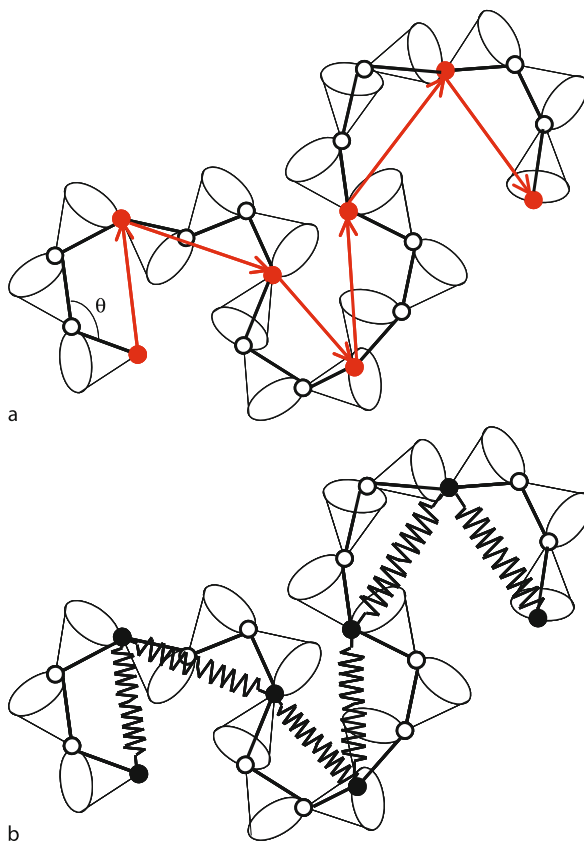
In this section we introduced typical topological indices which have a possibility to be suited to branched polymers. At present, the mathematical characterization of molecules has led to a large number of molecular descriptor, and their number continues to grow. These descriptors, often referred to as topological indices, will play an important role in structure-property and structure-activity studies. However, the important task is to find the physical meanings of these topological descriptors.

### Ideal Chain Models

In this section, we deal with only the statistical properties of ideal polymer chains in the unperturbed state and only short-range interactions are considered. Flory proposed that the polymer molecules in the molten state are unperturbed as they are in a  $\Theta$  solvent [42]. Small angle neutron scattering data [39,64] support the Flory theorem. An elegant explanation for the Flory theorem was given by de Gennes [29]. Therefore, the concept of ideal chain model plays a central role in analyzing the fundamental topological feature of any branched polymer.

Let us consider the simple homologous series of normal alkane hydrocarbons designated as linear polyethylenes. These polymers have a general structure where the number of  $-\text{CH}_2$ -units is connected by chemical single bonds. Although chemical bonds are fairly rigid with respect to stretching and to bending of the valence angles between adjacent bonds, a single-polymer molecule has many internal degrees of rotational freedom about each C-C bond in the polyethylene molecule, resulting in that it can adopt many different configurations, thereby necessitating the use of statistical mechanics.

Here we consider three models of varying levels of complexity and reality for flexible chains. The simplest model is called the "Kramers chain" [81] by considering the polymer to be composed of points or units that are freely joined by bonds of a fixed length. For the random flight chain the bond angles are fixed; however, there is free rotation about the bonds. This is called the "Kirkwood-Riseman chain" [76] (see Fig. 3a). The bond angle  $\theta$  of typical vinyl polymers is approximately  $111.5^\circ$  and the bond length is approximately 0.154 nm. More realistic models include the hindered rotations about the bonds that form the chain backbone, and they also possibly include the steric hindrances which result in the interdependence of the internal rotational degrees of freedom. Three types of micro-conformations, i. e. one trans and two gauche states, occur every C-C bond.



**Polymers, Non-linearity in, Figure 3**  
Kirkwood-Riseman chain and a its equivalent chain model; b its Gaussian model

Each conformation exists for only a very short time: the observed proportions of the microconformations, and thus the resulting macroconformations, are temporal averages over all molecules, resulting in that the polymer chain forms a random coil. Consider the end-to-end vector  $\mathbf{r}$  joining one end of the linear alkane chain to the other: the average length of the alkane chain can be considered as an indicator of the spreading out or the size of the polymer. If the chain is made up of bonds labeled from numeral 1 through  $\nu$ , atoms labeled from 1 through  $N$ , and if  $\alpha_i$  is the bond vector of the  $i$ th bond ( $i = 1, 2, 3, \dots, \nu$ ), we have

$$\mathbf{r} = \alpha_1 + \alpha_2 + \dots + \alpha_\nu. \quad (18)$$

It is apparent that we obtain the relation  $\nu = N - 1$  for any chain possessing no loops or rings. Since the average value of  $\mathbf{r}$  is zero, its average length can be obtained by taking the square root of  $r^2 = \mathbf{r} \cdot \mathbf{r}$ :

$$\langle r^2 \rangle_0 = \nu \alpha^2 + \left\langle 2 \sum_{i>j} \mathbf{r}_i \cdot \mathbf{r}_j \right\rangle, \quad (19)$$

where the angular bracket indicates the average of the spatial quantities, the subscript 0 signifies the  $\Theta$  conditions, and the double sum is over all pairs  $i < j$ . The end-to-end distance is the spatial size of a linear chain, but it has no significance for a branched chain. The square radius of gyration  $\langle s^2 \rangle$  and hydrodynamic Stokes radius  $r_H$  are an appropriate measure of the spatial size of linear and branched coils or particles [119,120].

The square radius of gyration is the means of all position of atoms from the center of mass vector. The mean square radius of gyration for unperturbed chains can be derived from the Lagrange theorem [30,43,59]:

$$\langle s^2 \rangle_0 = \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N \langle \mathbf{r}_{ij}^2 \rangle \quad (20)$$

and the hydrodynamic radius or the Stokes radius of the coil is obtained from the mean reciprocal distance:

$$\langle r_H^{-1} \rangle_0 = \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N \langle |\mathbf{r}_{ij}|^{-1} \rangle, \quad (21)$$

where  $\mathbf{r}_{ij}$  is the vector from atoms  $i$  to  $j$ .

In the random-flight Kramers model, the second term on the right-hand side of (19) disappears. Therefore we have

$$\langle r^2 \rangle_0 = \nu \alpha^2. \quad (22)$$

It follows that the size of the polymer is proportional to the square root of the bond number. The chain behaves as a ghost or phantom chain that can intersect or cross itself freely and is often called the “Markov chain”. Note that in the model  $\alpha$  is not always 0.154 nm, which is the actual length of the C-C bond. Mathematically, the possible maximum length of the freely joint chain is undoubtedly given by  $\nu \alpha$  which corresponds to the contour length  $r_{\text{cont}}$  because it paces off the contour of the chain. The contour length of a chain is given by geometry in an all-trans conformation with a fixed bond angle  $\theta$  as  $r_{\text{cont}} = n_C a \sin(\theta/2)$  where  $a$  is 0.154 nm and  $n_C$  is the number of C-C bonds. Thus the end-to-end distance of a fully extended real chain should be the maximum length,  $\nu \alpha$  of the straightly aligned Kramers chain. The value of  $\alpha$  may be the length of the virtual bond which is a projection of the real bonds onto the end-to-end vector. In this case, the virtual bond length is given by  $\alpha = 0.254$  nm and  $\nu = n_C/2$  because  $a = 0.154$  nm and  $\theta = 111.5^\circ$ .

The freely rotating chain is identical to the freely jointed chain in every respect, except that the angle between two adjoining bonds is held fixed at a predetermined value.

In this Karkwood–Risemann chain (see Fig. 3a), (19) becomes

$$\langle r^2 \rangle_0 = \left( \frac{1 - \cos \theta}{1 + \cos \theta} + \frac{2 \cos \theta [1 - (-\cos \theta)^\nu]}{\nu (1 + \cos \theta)^2} \right) \nu \alpha^2. \quad (23)$$

In a more realistic model, the symmetric hindered rotation model, the conformations of the bonds do not occur with equal probability; however, they are determined by a torsion angle-dependent potential function  $u(\phi)$ , where  $\phi = 0^\circ$  for the *trans* state and  $\phi = \pm 120^\circ$  for the *gauche* state. At equilibrium, the bond conformations are distributed according to the Boltzmann distribution, and we obtain

$$\langle r^2 \rangle_0 = \left( \frac{1 - \cos \theta}{1 + \cos \theta} + \frac{2 \cos \theta [1 - (-\cos \theta)^\nu]}{\nu (1 + \cos \theta)^2} \right) \times \left( \frac{1 - \langle \cos \phi \rangle}{1 + \langle \cos \phi \rangle} \right) \nu \alpha^2, \quad (24)$$

where the average value of  $\langle \cos \phi \rangle$  can be readily obtained from the assumed potential energy function.

Consequently both these realistic models in the unperturbed state adapt to the same type of equation as

$$\langle r^2 \rangle_0 = C \nu \alpha^2, \quad (25)$$

where  $C$  is called the characteristic ratio and depends on the nature of the polymer. As is evident from (23) and (24), the  $C$  is slightly dependent on the value  $\nu$  but  $C$  may be considered to be constant for flexible polymers or large  $\nu$ . The chain composed of  $\nu$  bonds of length  $\alpha$  can be regarded as a freely jointed chain consisting of  $n$  “effective bonds” of length  $b$  by taking  $b = C\alpha$  and  $n = \nu/C$ . The two parameters  $n$  and  $b$  are determined under the restriction that the contour length must be the same as that of the real chain, i. e.,  $\nu \alpha = nb$ . Then, the end-to-end distance of any unperturbed chain molecule can be renormalized as

$$\langle r^2 \rangle_0 = nb^2. \quad (26)$$

Such a coarse-graining polymer chain is often called the *equivalent chain*. The effects of the chemical bulkiness of units, the inclusions of double bonds and short chain branches can be attributed to the parameter  $C$ . Thus, a real flexible polymer may be effectively treated as a freely jointed chain; i. e. Kramers chain, of  $n$  bonds with a length of  $b$ , and the details of the chemical structure of real chains can be smeared out (see Fig. 3a).

The equilibrium distribution function in freely jointed chains is assumed to be identical to the random walk. It should be noted here that the random walk distribution

is inconsistent with the equilibrium statistical mechanics [52] but the differences are probably inconsequential for large  $N$ . As described previously, a polymer chain can take many different macroconformations above a molecular glass temperature. This corresponds to the micro-Brownian motion and the chains are rapidly converted into other macroconformations. The instantaneous shape, which is obtained by time-averaging over many conformations, can be described by a Gaussian distribution for large  $N$ . The distribution function  $W(\mathbf{b}_i)$  of the vector  $\mathbf{b}_i$  between adjacent units  $i$  and  $i + 1$  can be defined either as the time-averaged incidence of  $\mathbf{b}_i$  within the specified range for a given molecule or as the average incidence for an ensemble of many identical units subject to identical conditions:

$$W(\mathbf{b}_i) = \left( \frac{3}{2\pi \langle \mathbf{b}^2 \rangle} \right)^{3/2} \exp \left( -\frac{3\mathbf{b}_i^2}{2\langle \mathbf{b}^2 \rangle} \right), \quad (27)$$

where  $\langle \mathbf{b}^2 \rangle$  is the time-averaged square bond length. Let the position vector of the  $i$ th effective unit be  $\mathbf{r}_i$ . Then, the distribution of the effective bond vector  $\mathbf{b}_i = \mathbf{r}_{i+1} - \mathbf{r}_i$  is given by a Gaussian distribution function (27); hence the probability distribution of the set of position vectors  $\{\mathbf{r}_i\} = \{\mathbf{r}_1, \dots, \mathbf{r}_N\}$  is proportional to

$$P(\{\mathbf{r}_i\}) = \left( \frac{3}{2\pi \langle \mathbf{b}^2 \rangle} \right)^{3/2} \exp \left( -\frac{3}{2\langle \mathbf{b}^2 \rangle} \sum_{i=1}^{N-1} (\mathbf{r}_{i+1} - \mathbf{r}_i)^2 \right). \quad (28)$$

The equilibrium state of this chain is described by a distribution function proportional to  $\exp(-V/k_B T)$ , where  $V$  is the potential energy,  $k_B$  is the Boltzmann constant and  $T$  is the absolute temperature; therefore, if we choose

$$V = \frac{1}{2} k \sum_{i=1}^{N-1} (\mathbf{r}_{i+1} - \mathbf{r}_i)^2, \quad (29)$$

where  $k = 3k_B T / \langle \mathbf{b}^2 \rangle$ , then the chain's equilibrium distribution is given by (28). Consequently, the Gaussian chain can be modeled as a chain of beads connected by a Hookean spring with a spring constant  $k$  of (29) as shown in Fig. 3b. In the model, if the square root of  $\langle \mathbf{b}^2 \rangle$  is identical to the effective bond length  $b$ ,  $\langle \mathbf{b}^2 \rangle^{1/2} = b$ , then the end-to-end distance of Gaussian chain may be identical to the value given by (26). In this case, the spring of the Gaussian chain is called the "segment" (see Fig. 3b). The Gaussian chain plays a central role in the study of chain dynamics [17,110,137]. The applicability of the Gaussian chain model to any type of branched polymers was soon

attempted [22,23,32,67,79,138]. The length of any bond and the angles between the bonds are insignificant because a connection between segments, rather than the precise nature of the connection of atoms, is important in topological analysis for flexible polymers.

### Chain Statistics of Nonlinear Polymers

One of the quantitative measures that reflect molecular branching is the average size of isolated single branched polymers. The expansion in the space of isolated branched polymers is smaller than that of linear chains with the same molecular weight. As shown in the previous section, we can treat any flexible polymer as a random-flight chain model which can be regarded as a molecular graph composed of  $N - 1$  statistical bonds (edges) of a length  $b$  joining  $N$  points (vertices) of unit mass. The quality  $b$  is the segment length and it depends on the conformational characteristics of the polymer species.

When the branched chains obey random-flight statistics, the average of the scalar product extending over all sets of bond vectors  $\mathbf{b}_i$  is given by  $\langle \mathbf{b}_i \cdot \mathbf{b}_j \rangle = b^2 \delta_{ij}$  where  $\delta_{ij}$  is the Kronecker delta function. Thus, the mean square of  $\mathbf{r}_{ij}$  becomes

$$\langle \mathbf{r}_{ij}^2 \rangle = \sum_{i,j} b^2 = d_{ij} b^2, \quad (30)$$

because the number of bonds between the  $i$ th and  $j$ th bead is identical to the graph-theoretical distance  $d_{ij}$ . Then, combination of (13) and (20) with (30) yields

$$\langle s^2 \rangle_0 = \left( \frac{b}{N} \right)^2 W. \quad (31)$$

This is a graph-theoretical expression for the mean square radius of gyration for any unperturbed random-flight chain [96,132]. The radius of branched chains becomes proportional to  $W^{1/2} N^{-1}$ .

As a measure of branching, Zimm-Stockmayer proposed a parameter  $g$  [139] which is defined by the ratio of  $\langle s^2 \rangle$  for a given branched chain to that for the linear chain at the same number of statistical units. Consequently, the  $g$ -factor is equal to the Wiener index normalized by the Wiener index of linear chain proposed by Bonchev et al. [9,10]:

$$g = \frac{\langle s^2 \rangle_0}{\langle s^2 \rangle_0^L} = \frac{W}{W_L}, \quad (32)$$

where the superscript L indicates the linear chain. The Wiener index of linear chains  $W_L$  can be easily evaluated



by substituting the relation  $k_i = N - i$  into (12):

$$W_L = \sum_{i=1}^{N-1} i(N-i) = \frac{1}{6} N(N^2 - 1). \quad (33)$$

Substituting (33) into Eq. (31) yields the equation of  $\langle s^2 \rangle$  of linear chains, which was first derived by Kramers [31,81,112]. Substitution of (33) into (32) results in that the  $g$ -factor of any branched polymer can be represented by the Wiener index  $W$  and the number of beads  $N$ :

$$g = 6 \frac{W}{N(N^2 - 1)} \cong 6N^{-3} W. \quad (34)$$

Here we present the graph theoretical expressions of typical types of branched random-flight chains to evaluate their mean square radius of gyration according to the mathematical procedure by Bonchev-Trinajstić [15]. The branched chains are here classified into two types: one is a type of a non-linear chain with a fixed type of branching: referred to as a “specifically branched chain”, and the other is an unfixed type of branching: referred to as a “randomly branched chain”.

### Specifically Branched Chains

Let us consider a chain with any specified branching geometry in which various linear side chains are arbitrarily connected with a linear main chain. Bonchev-Trinajstić [15] showed that the Wiener index is given by the sum of the total path number between any two beads in the main chain,  $W_{\text{ch}}$ , that between any two beads in a common side chain,  $W_{\text{br}}$ , that between a bead in the main chain and a bead in a side chain,  $W_{\text{ch-br}}$  and that between a bead in a side chain and a bead in another side chain,  $W_{\text{br-br}}$ .

The molecular graph of such a specifically branched chain is illustrated in Fig. 4, in which  $N_0$  is the total number of beads in the main chain and the beads are labeled with the numerals 1 through  $N_0$ ;  $p$  is the total number of branched beads;  $m_k$  is the label numeral of the  $k$ th branched bead, where  $k$  enumerates the branched beads ( $1 \leq k \leq p$ );  $n_k$  is the total number of branches connected with the labeled bead  $m_k$ ; and  $N_{k,j}$  is the number of beads in the  $j$ th branch, where  $j$  enumerates the branches attached to the bead  $m_k$ . Then, the Wiener index of graphs with linear branches is given by

$$W = W_{\text{ch}} + W_{\text{br}} + W_{\text{ch-br}} + W_{\text{br-br}}, \quad (35)$$

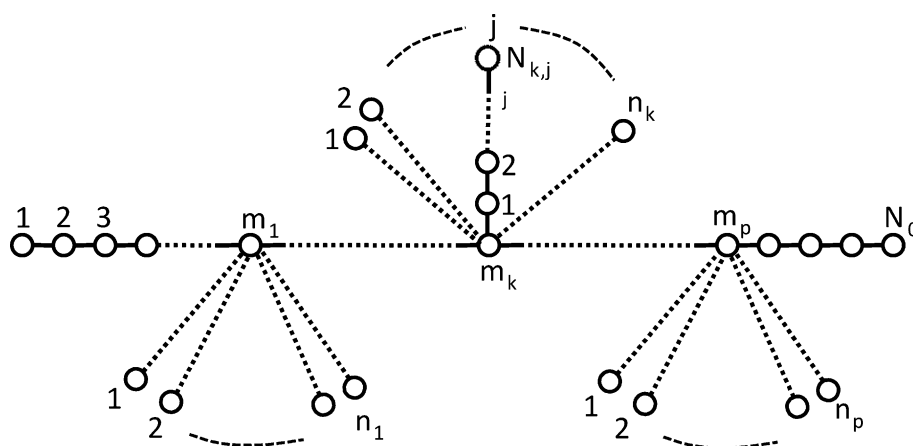
where

$$W_{\text{ch}} = \frac{1}{6} N_0 (N_0^2 - 1), \quad (36)$$

$$W_{\text{br}} = \sum_{k=1}^p \left[ \sum_{j=1}^{n_k} \frac{1}{6} N_{k,j} (N_{k,j}^2 - 1) \right], \quad (37)$$

$$W_{\text{ch-br}} = \sum_{k=1}^p \sum_{j=1}^{n_k} \frac{1}{2} N_{k,j} \left[ (N_0 - m_k + 1)^2 + (m_k^2 - 1) + N_0 N_{k,j} \right], \quad (38)$$

$$W_{\text{br-br}} = \sum_{k=1}^p \left[ \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} N_{k,i} N_{k,j} \times \left( \frac{N_{k,i} + N_{k,j}}{2} + 1 \right) \right]$$



Polymers, Non-linearity in, Figure 4

The molecular graph of a specifically branched chain with linear branches

$$+ \sum_{k=1}^{p-1} \sum_{l=1+k}^p \left[ \sum_{i=1}^{n_l} \sum_{j=1}^{n_k} N_{l,i} N_{k,j} \times \left( \frac{N_{l,i} + N_{k,j}}{2} + 1 + m_l - m_k \right) \right]. \quad (39)$$

Here the total number of beads  $N$  becomes

$$N = N_0 + \sum_{k=1}^p \sum_{i=1}^{n_k} N_{k,i}. \quad (40)$$

The first term on the right-hand side of (39) is the sum of the path number extending over all pairs of branches connected with a common branched bead and the second term is the sum extending over all pairs of branches connected with different branched beads. Thus, the first term vanishes for tri-functional comb-shaped molecules and the second term vanishes for star-shaped molecules.

Explicit expressions for the mean square radius of gyration  $\langle s^2 \rangle_0$  or the  $g$ -factor can be obtained by the substitution of (35)–(39) into (31) or (34). Furthermore, the present formula from (36) to (39) for specifically branched chains with linear branches can be recursively extended to include more complicated types of specifically branched chains having any branched branches as well as reduced to simplified types of specifically branched chains. The changes in the Wiener index due to the linkage of two graphs are formulated in general by Polansky et al. [105,106].

Here we exemplify the expressions for the Wiener indices of two specifically branched chains that are of great interest in practical use, that is, star-shaped and tri-functional comb-shaped chains.

Setting  $p = 1$  in from (36) to (39), we obtain the Wiener index of stars:

$$\begin{aligned} W_{\text{Star}} = & \frac{1}{6} N_0 (N_0^2 - 1) + \sum_{j=1}^{n_1} \frac{1}{6} N_{1,j} (N_{1,j}^2 - 1) \\ & + \frac{1}{2} \sum_{j=1}^{n_1} N_{1,j} \left[ (N_0 + 2 - m_1) (N_0 + 1 - m_1) \right. \\ & \quad \left. + (m_1 - 1) (m_1 + 2) + N_0 (N_{1,j} - 1) \right] \\ & + \sum_{i=1}^{n_1-1} \sum_{j=1+i}^{n_1} N_{1,i} N_{1,j} \left( \frac{N_{1,i} + N_{1,j}}{2} + 1 \right). \end{aligned} \quad (41)$$

Likewise, setting  $n_k = 1$  in (36)–(39), we obtain that of combs:

$$\begin{aligned} W_{\text{Comb}} = & \frac{1}{6} N_0 (N_0^2 - 1) + \sum_{k=1}^p \frac{1}{6} N_{k,1} (N_{k,1}^2 - 1) \\ & + \frac{1}{2} \sum_{k=1}^p N_{k,1} \left[ (N_0 + 2 - m_k) (N_0 + 1 - m_k) \right. \\ & \quad \left. + (m_k - 1) (m_k + 2) + N_0 (N_{k,1} - 1) \right] \\ & + \sum_{k=1}^{p-1} \sum_{j=1+k}^p N_{l,1} N_{k,1} \left( \frac{N_{l,1} + N_{k,1}}{2} \right. \\ & \quad \left. + 1 + m_l - m_k \right). \end{aligned} \quad (42)$$

The topological method gives the results numerically identical with that calculated by usual statistical methods [5,99]. Advantages in the formulation based on the Wiener index are that the mean square radius of gyration or the  $g$ -factor not only can be evaluated in the homologous process but also can be expressed as a function of topological parameters such as the position of branches and the length of main or side chains. For example, let us consider a simple specifically branched chain with one linear side chain, the so-called Y-shaped chain. The Wiener index of the Y-shaped chain  $W_Y$  can be readily evaluated from (41) or (42). The  $g$ -factor of Y-shaped chain  $g_Y$  can be expressed as a function of the position of branch point or the length of branches.

$$g_Y = \frac{6N_{1,1}}{N(N^2 - 1)} \left[ (m_1 - 1)^2 - (N_0 - 1)(m_1 - 1) \right] + 1, \quad (43)$$

where  $N_{1,1}$  is the number of beads in the branch,  $N_0$  is the number of beads in the main chain,  $N = N_{1,1} + N_0$ , and  $m_1$  is the position of branched bead on the main chain. When  $N_{1,1}$  and  $N_0$  are fixed and discrete,  $m_1$  is transformed into continuous  $m_1$  because of the large  $N_0$  and  $g$  becomes a quadratic function of the branch position  $m_1$  in which  $g$  has the minimum at  $m_1 = N_0/2$ . Likewise, when  $N$  and  $m_1$  are fixed,  $g$  becomes another quadratic function of  $N_0$  or  $N_{1,1}$  and has a minimum at  $N_0 = (N + m_1)/2$  or  $N_{1,1} = (N - m_1)/2$ .

Block copolymers [126] have also been given much attention to offer the potential for obtaining materials that incorporate the properties of different homopolymers and for improving the compatibility of the immiscible polymer blends and the interfacial adhesion in polymeric composites. Recent techniques of living anionic copolymer-

ization [66] make it possible to produce various complicated nonlinear copolymers such as graft copolymers with the different chemical nature of the backbone [117,118] and the attached branches and the miktoarm ( $\mu$ -star), or heteroarm star polymers consisting of chemically different arms [65]. The graph-theoretical expressions for linear block copolymers has been studied by Yang [136].

The nonlinear copolymer thus considered is a chain consisting of branch units with edge bond length  $b$  attached to the backbone chain with edge length  $a$ . The length of edge of backbone and branches is different so that the path number can be determined on the basis of the concept of Altenburg polynomial [1]. The path number of the nonlinear graph can be obtained

$$\Omega_{\text{ch}} = \frac{1}{6} N_0 (N_0^2 - 1) a^2, \quad (44)$$

$$\Omega_{\text{br}} = \sum_{k=1}^p \left[ \sum_{j=1}^{n_k} \frac{1}{6} N_{k,j} (N_{k,j}^2 - 1) b^2 \right], \quad (45)$$

$$\Omega_{\text{ch-br}} = \sum_{k=1}^p \sum_{j=1}^{n_k} \frac{1}{2} N_{k,j} \left[ \left\{ (N_0 - m_k + 1)^2 + (m_k^2 - 1) \right\} a^2 + N_0 N_{k,i} b^2 \right], \quad (46)$$

$$\begin{aligned} \Omega_{\text{br-br}} = & \sum_{k=1}^p \left[ \sum_{i=1}^{n_k-1} \sum_{j=1+i}^{n_k} N_{k,i} N_{k,j} \left( \frac{N_{k,i} + N_{k,j}}{2} + 1 \right) b^2 \right. \\ & \left. + a^2 \right] + \sum_{k=1}^{p-1} \sum_{l=1+k}^p \left[ \sum_{i=1}^{n_l} \sum_{j=1}^{n_k} N_{l,i} N_{k,j} \right. \\ & \left. \times \left\{ \left( \frac{N_{l,i} + N_{k,j}}{2} \right) b^2 + (1 + m_l - m_k) a^2 \right\} \right]. \end{aligned} \quad (47)$$

The total path number statistically weighted by  $a$  and  $b$  becomes  $\Omega = \Omega_{\text{ch}} + \Omega_{\text{br}} + \Omega_{\text{ch-br}} + \Omega_{\text{br-br}}$  and the mean square radius of gyration for any nonlinear copolymers is given by  $\langle s^2 \rangle = \Omega/N^2$ . Of course (44)–(47) of  $\Omega$  are reduced to (36)–(39) of the Wiener index  $W$  for branched homopolymers when  $a = b = 1$ .

### Randomly Branched Chains

Any real polymeric material will have a distribution in chain lengths, number and position of branches, and so on, depending on polymerization and/or fractionation; the actual polymers may be a mixture of such isomers. For the proceeding to such actual polymers, Zimm–

Stockmayer [139] investigated the ensemble of random mixture of structural isomers, in which it was assumed that the total number of beads (or molecular weight) and the number of subchains per chain are fixed but all possible arrangements of subchains of various lengths occur with equal frequency. Here the term “subchain” refers to a portion of the molecule between two adjacent branched beads or between adjacent pair of an end and a branched bead.

An example of a randomly branched chain is shown in Fig. 5a. Let  $n_k$  be the number of the beads within the  $k$ th subchain, then the total number of beads is given by  $N = n_1 + n_2 + \dots + n_p$  where  $p$  is the total number of subchains per molecule. Let us consider the  $g$ -factor of randomly branched chains with a random distribution of  $n_k$  which is defined such that all possible sets of the numbers,  $(n_1, n_2, \dots, n_p)$ , occur with equal frequency under the restriction that  $N$  is fixed. Then the  $g$ -factor of the randomly branched chains can be expressed as a function of a set of  $(n_1, n_2, \dots, n_p)$ . Consequently the  $g$ -factor of any branched chain can be determined by averaging the  $g$  value over all arrangements of  $n_k$  under the fixed  $N$  and  $p$  as follows

$$\begin{aligned} \bar{g} &= \text{Av}[g] \\ &= \frac{\int_0^N dn_{p-1} \int_0^{N-n_{p-1}} dn_{p-2} \dots \int_0^{N-\sum_{k=1}^{p-1} n_k} dn_1 g(n_1, n_2, \dots, n_p)}{\int_0^N dn_{p-1} \int_0^{N-n_{p-1}} dn_{p-2} \dots \int_0^{N-\sum_{k=1}^{p-1} n_k} dn_1}. \end{aligned} \quad (48)$$

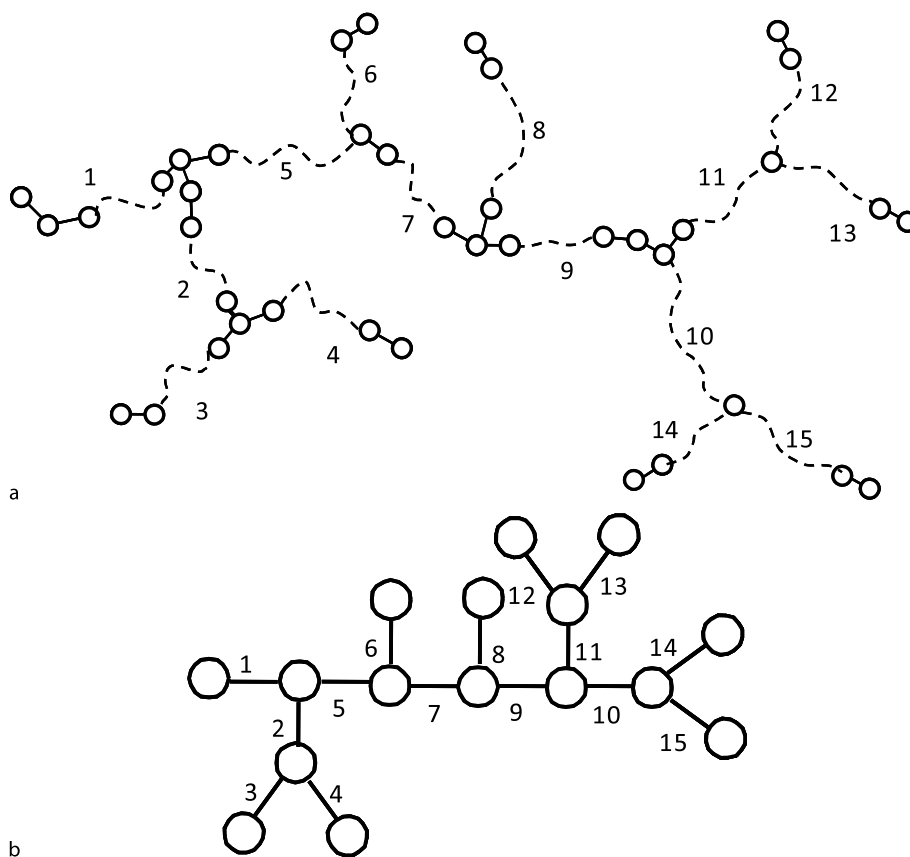
According to Kataoka and Saito [83,135], the  $g$ -factor can be expressed using the total number of subchains  $p$  as

$$\bar{g} = \frac{6}{p(p+1)(p+2)} \left( p^2 + \sum_{(\alpha,\beta)} v_{\alpha\beta} \right), \quad (49)$$

where  $v_{\alpha\beta}$  represents the number of subchains lying between the specified subchains  $\alpha$  and  $\beta$ , and  $\sum_{(\alpha,\beta)}$  is the sum extending over all distinguishable pairs of subchains.

According to Nitta [96], the  $g$ -factor of any randomly branched chain can be estimated from the Wiener index  $\omega$  of its reduced graph in which subchains are transformed into single edges of unit length and both junctions and ends into vertices of unit mass, as shown in Fig. 5b. Thus,

$$\bar{g} = \text{Av}[g] = \text{Av} \left[ \frac{W}{W_L} \right] = \frac{\omega}{\omega_L}, \quad (50)$$



Polymers, Non-linearity in, Figure 5

A randomly branched molecule and its reduced molecular graph in which branched and end beads are transformed to uniform vertices and subchains are transformed to uniform edges

where  $\omega_L$  is the Wiener index of a linear graph with the same number of edges which can be readily evaluated from (33) by setting  $N = p + 1$ . Thus,

$$\omega_L = \frac{p(p+1)(p+2)}{6}. \quad (51)$$

Moreover, since  $W_L$  may be considered to be constant for the averaging process in (48), the average Wiener index can be readily obtained as follows:

$$\bar{W} = \text{Av}[W] = \text{Av}[g] W_L = \frac{N(N+1)(N+2)}{p(p+1)(p+2)} \omega. \quad (52)$$

Consequently, the mean Wiener index of any randomly branched chain can be calculated by the total number of subchains  $p$ , the total number of beads  $N$ , and the Wiener index  $\omega$  of the corresponding reduced graph where all subchains are transformed into single edges (see Fig. 5b). This indicates that the mathematical difficulty involved with the treatment of highly branched molecules is largely reduced

by the use of the Wiener index. The reduced graph corresponds to the *proper graph*, which contains only two types of vertices-terminal and branched ones, as introduced by Bonchev et al. [10]. Specific cases of randomly branched chains are exemplified below.

In the case of “star” chains, the Wiener index can be evaluated from (41) by putting  $N_{1,i} = 1$ ,  $N_0 = 3$ ,  $m_1 = 2$  and  $n_1 = q - 2$ :

$$\omega_S = q^2. \quad (53)$$

In the case of the “comb” type chains, we obtain

$$\omega_C = \frac{1}{12} (q^3 + 9q^2 - q + 3) \quad (54)$$

by substituting the relations that  $N_{k,1} = 1$ ,  $m_k = k + 1$ ,  $p = (q - 1)/2$ , and  $N_0 = (q + 3)/2$  into (42). Of course, these equations are identical with the results derived from Kurata and Fukatsu’s statistical treatments [83]. Gutman et al. [62,63] presented the formula for the Wiener index of

regular dendrimer graphs of three and four degrees. The more generalized Wiener index for a regular dendrimer graph of degree  $f$  is given by

$$\omega_D(n, f) = \frac{f}{(f-2)^3} \left[ \{nf^2 - 2(1+n)f + 1\} (f-1)^{2n} + 2f(f-1)^n - 1 \right], \quad (55)$$

where  $n$  is the generation number.

### Intrinsic Viscosity and Hydrodynamic Radius

The most precise and direct experimental method for characterizing chain architecture is through the measurement of dilute solution properties. The most useful measure of branching is the Zimm–Stockmayer ratio  $g$  of the radius of gyration of the branched polymer to that of the linear polymer at the same molecular weight as shown in (32). Other commonly employed branching parameters are the ratio of the intrinsic viscosity  $g'$  [138] and the ratio of hydrodynamic radius  $h$  [120] for branched polymers to the linear ones, respectively. The non-free draining intrinsic viscosity of suspended spheres with hydrodynamic radius was given by Einstein as

$$[\eta]_{ND} = \frac{5}{2} \left( \frac{N_A}{M} \right) \frac{4}{3} \pi r_H^3, \quad (56)$$

where  $M$  is the molecular weight and  $N_A$  is Avogadro's number.

According to (21), the mean reciprocal hydrodynamic radius  $r_H^{-1}$  can be calculated from the mean reciprocal of  $\mathbf{r}_{ij}$  given by [59,120]

$$\langle |\mathbf{r}_{ij}|^{-1} \rangle = \left( \frac{6}{\pi} \right)^{1/2} d_{ij}^{-1/2} b^{-1}. \quad (57)$$

Using the generalized Harary index (16) with  $\varepsilon = 1/2$ , one obtains the reciprocal hydrodynamic radius given by

$$\langle r_H^{-1} \rangle_0 = \frac{1}{bN^2} \left( \frac{6}{\pi} \right)^{1/2} H_{1/2} \quad (58)$$

and its branching factor can then be written as

$$h = \frac{\langle r_H \rangle_0^{-1}}{\langle r_H \rangle_0^{L-1}} = \frac{H_{1/2}^L}{H_{1/2}}, \quad (59)$$

where the superscript L represents the linear chain. In the non-draining condition, we have  $g' = h^3$ . Consequently the hydrodynamic radius and the branching parameter  $h$  or  $g'$  can be evaluated from the distance matrix.

Assuming that the Stokes radius  $r_H$  is proportional to the square root of the mean-square radius of gyration with proportional constant  $\chi$

$$r_H = \chi (6 \langle s^2 \rangle)^{1/2} = \chi \left( \frac{b}{N} \right) (6W)^{1/2}. \quad (60)$$

Consequently the intrinsic viscosity in the non-draining condition is also related to the Wiener index. Consequently we have [44]

$$[\eta]_{ND} = \frac{5}{2} \Phi \frac{(6 \langle s^2 \rangle)^{3/2}}{N} = \frac{5}{2} \Phi \frac{b^3}{N^4} (6W)^{3/2}, \quad (61)$$

where  $\Phi = (N_A/M_0) 4/3 \pi \chi^3$  and  $M_0$  is the molar mass of the unit, i. e.  $M_0 = M/N$ .

### Chain Dynamics of Nonlinear Polymers

As mentioned before, Gaussian chains have played a central role on the statistics and dynamics of flexible chain molecules with and without branching [8,53,78,100]. According to the central limit theorem in statistical physics [59], the random-flight statistics of a flexible polymer can be described by a Gaussian chain which is mathematically simpler to handle [47]. Rouse [110] and Bueche [17] demonstrated that the dynamics of dilute solutions of linear polymers can be characterized by considering a Gaussian chain suspended in a flowing viscous liquid. Subsequently, the extension of this model to any branched molecule was made by Ham [67]. In fact, the Rouse model provides a good description for semi-dilute solutions and for melts below the entanglement limit rather than dilute solutions [38]. The mathematical representation of graph theory is helpful to generalize the statistics and dynamics of Gaussian chains to include any type of branching [36,45,46,136]. In this section, we present a graph-theoretical method for calculating the relaxation spectra of flexible chain molecules with any type of branching. The significance of this approach lies in providing the relationship between the statistics and the dynamics of chain molecules being reformulated into a more convenient algebraic form.

#### Rouse–Ham Theory

Branched Gaussian chains containing no loops or circles contains  $N$  beads and  $N - 1$  segments acting as Hookean springs with spring constant of  $3k_B T/b^2$  as shown in (29). The laws, which govern the behavior of linear flexible chain, are assumed to hold for nonlinear flexible polymers. Considering a Gaussian chain suspended in a flow liquid,



one finds the motion equation given by [101]

$$-\zeta_0 \dot{\mathbf{r}} = \frac{3k_B T}{b^2} \mathbf{Z} \cdot \mathbf{r}, \quad (62)$$

where  $\mathbf{r}$  is a  $3 \times N$  matrix whose rows contain the dimensional component of the  $N$  position vectors of beads,  $\dot{\mathbf{r}}$  is the time derivative of  $\mathbf{r}$ ,  $\zeta_0$  is the friction constant of the beads, and  $\mathbf{Z}$  is the  $N \times N$  connectivity matrix which is called the Zimm matrix in polymer physics [137] and is identical to the Laplacian matrix  $\mathbf{L} (= \mathbf{V} - \mathbf{A})$  in graph theory. According to Rouse theory [110], one can rewrite (62) in terms of bond vectors  $\mathbf{b}$  as

$$-\zeta_0 \dot{\mathbf{b}} = \frac{3k_B T}{b^2} \mathbf{R} \cdot \mathbf{b}, \quad (63)$$

where  $\mathbf{b}$  is a  $3 \times (N - 1)$  matrix whose rows contain the dimensional component of the  $N - 1$  bond vectors. The matrix  $\mathbf{R}$  is the  $(N - 1) \times (N - 1)$  connectivity matrix which is called the Rouse matrix [110]. Comparing the characteristic polynomials of  $\mathbf{Z}$  and  $\mathbf{R}$ , we can find the following relation:

$$\Phi(\mathbf{Z}; \lambda) = \lambda \Phi(\mathbf{R}; \lambda). \quad (64)$$

The eigenvalues of  $\mathbf{Z}$  contain one zero eigenvalue and, hence,  $\mathbf{Z}$  do not possess an ordinary inverse. (64) shows that the non-zero eigenvalues of  $\mathbf{Z}$  are identical with those of  $\mathbf{R}$ . The springs and beads are assigned in any arbitrary fashion. The eigenvalues of  $\mathbf{Z}$  and  $\mathbf{R}$  are, however, independent of how their elements are numbered [46,51,94]. Also, this relation (64) corresponds to the relation (11).

The zero eigenvalue of  $\mathbf{Z}$  represents the mode of chain translation [61]. Each eigenvalue  $\lambda_i$  ( $i = 1, 2, \dots, N - 1$ ) of  $\mathbf{R}$  or non-zero eigenvalue of  $\mathbf{Z}$  is associated with the relaxation times  $\tau_i$  of the  $i$ th mode for dynamic molecular motions [46,51]:

$$2\tau_i = \frac{\zeta_0 b^2}{3k_B T} \lambda_i^{-1}. \quad (65)$$

According to the theory of linear viscoelasticity, the relaxation time spectrum,  $H(\tau)$ , is given by

$$H(\tau) = \frac{ck_B T}{N} \sum_{i=1}^{N-1} \delta(\ln \tau - \ln \tau_i), \quad (66)$$

where  $c$  is the concentration of beads per unit volume and  $\delta(x)$  is a Dirac delta function. Various rheological functions describing the bulk properties of polymers can be determined using the relaxation spectrum  $H(\tau)$ .

The Zimm matrix  $\mathbf{Z}$  or the Rouse matrix  $\mathbf{R}$  can also be constructed in a different manner by making use of the incident matrix of a digraph  $D$ . Forsman [46] showed that  $\mathbf{Z}$  and  $\mathbf{R}$  are given by the incidence matrix  $\mathbf{B}$  of  $D$ :

$$\mathbf{Z} = \mathbf{B}\mathbf{B}^\top \quad \text{and} \quad \mathbf{R} = \mathbf{B}^\top \mathbf{B}. \quad (67)$$

It should be here noted that all elements of a connectivity matrix  $\mathbf{K}$  formed by  $\mathbf{C}^\top \mathbf{C}$  are the absolute values of those of  $\mathbf{R}$ . Combination of the relations (10), (11), (64),  $\mathbf{Z} = \mathbf{L}$  gives

$$\Phi(\mathbf{R}; \lambda) = \Phi(\mathbf{K}; \lambda). \quad (68)$$

Furthermore from the Kirchhoff's relation (4), we have

$$\Phi(\mathbf{R}; \lambda) = \Phi(\mathbf{A}_L; \lambda - 2). \quad (69)$$

Consequently, the eigenvalues  $\lambda_i$  of  $\mathbf{R}$  can be calculated from the eigenvalues  $\mu_i$  of  $\mathbf{A}_L$  as follows:

$$\lambda_i = \mu_i + 2. \quad (70)$$

Remembering (65), we have

$$\tau_i = \frac{\zeta_0 b^2}{6k_B T} (\mu_i + 2)^{-1}. \quad (71)$$

It is interesting to note that the relaxation spectrum of a chain molecule can be determined entirely by the set of eigenvalues of the adjacency matrix  $\mathbf{A}_L$  of its line graph.

According to Forsman [46], the square radius of gyration is given by  $s^2 = N^{-1} \mathbf{b}^\top \cdot \mathbf{F} \cdot \mathbf{b}$  where  $\mathbf{F}$  is called the Kramers matrix and it is equal to the inverse of  $\mathbf{R}$ . When the chains obey random-flight statistics, the mean square of radius of gyration can be related to the reciprocal of eigenvalues of  $\mathbf{R}$  and we have

$$\langle s^2 \rangle_0 = \frac{b^2}{N} \text{Tr}(\mathbf{R}^{-1}) = \frac{b^2}{N} \sum_{i=1}^{N-1} \frac{1}{\lambda_i} = \frac{b^2}{N} \sum_{i=1}^{N-1} \frac{1}{\mu_i + 2}, \quad (72)$$

where  $\text{Tr}$  denotes the trace of a matrix. The mean radius of gyration is related to the sum of the reciprocal eigenvalues of Rouse matrix so that it largely depends on the minimum eigenvalue of  $\mathbf{R}$  or the second smallest one of  $\mathbf{L}$ . This corresponds to Mohar's suggestion that the second smallest one of  $\mathbf{L}$  is related to the diameter and mean distance in a graph [94].

Comparing (72) and (31) gives

$$W = N \text{Tr}(\mathbf{R}^{-1}) = N \sum_{i=1}^{N-1} \frac{1}{\lambda_i} = N \sum_{i=1}^{N-1} \frac{1}{\mu_i + 2}. \quad (73)$$

Moreover, a combination of (65) and (73) gives a relation between the Wiener index and relaxation times:

$$W = \frac{6Nk_B T}{\xi b^2} \sum_{i=1}^{N-1} \tau_i. \quad (74)$$

This may be a physical meaning of the Wiener index. The steady-state viscosity in a dilute solution can be obtained from the relaxation time spectrum (66):

$$\eta = \eta_s + c \frac{N_A}{M} k_B T \sum_{i=1}^{N-1} \tau_i. \quad (75)$$

Consequently, the intrinsic viscosity in the freely draining condition can be written as

$$[\eta]_{\text{FD}} = \frac{N_A}{M\eta_s} k_B T \sum_{i=1}^{N-1} \tau_i, \quad (76)$$

which turns out to be proportional to the mean square of the radius of gyration:

$$[\eta]_{\text{FD}} = \frac{N_A \xi}{6M_0 \eta_s} \langle s^2 \rangle = \frac{N_A \xi b^2}{6M_0 \eta_s N^2} W, \quad (77)$$

where  $\eta_s$  is the viscosity of pure solvent. It can be shown that the intrinsic viscosity in freely draining conditions is proportional to  $W/N^2$  for branched chains and to the first power of  $N$  for linear chains. (77) may be referred to as “Staudinger’s relation”. This suggests that the Wiener index of any polymer molecule can be evaluated from the steady-flow viscosity of dilute solution in the freely draining condition.

The non free-draining intrinsic viscosity  $\eta_{\text{ND}}$  as shown in (61) is proportional to  $W^{3/2}/N^4$ . Zimm and Kilb [138] introduced a branching parameter  $g'$  defined as the ratio of  $[\eta]$  of a branched polymer to that of a linear polymer in the  $\Theta$  condition. Therefore, we have  $g' = g^{1.0}$  for the freely draining condition and  $g' = g^{1.5}$  for the non-draining condition. For various branched polymers, however, measured exponents fall between 0.5 and 2.0, and they are strongly dependent on solvents and molecular weight [82,125].

Sheridan et al. [114] examined the relation between the intrinsic viscosity and the Wiener index for hyper-branched polymers. According to their computer simulations, they found the relation  $[\eta] \sim W^a N^b$  where  $a = 1.0$  and  $b = -2.2$  which is similar to the result of (77) in the freely draining condition.

## High Moments of Relaxation Time and Radius of Gyration

The potential energy of the chain molecule given by (29) can be rewritten as

$$V = \frac{1}{2} k \text{Tr} [\mathbf{r} \cdot \mathbf{r}^\top]. \quad (78)$$

Fixman [40] derived that the distribution function of the square radius of gyration  $s^2$  as follows:

$$P(s^2) \propto \int \delta(s^2 - N^{-1} \text{Tr} [\mathbf{r} \cdot \mathbf{r}^\top]) e^{-V/k_B T}. \quad (79)$$

In order to obtain the higher moments of radius of gyration  $\langle s^{2n} \rangle$ , we introduce a Laplace transform of  $P(s^2)$  with respect to  $s^2$  according to Eichinger’s treatment [35]. Then we obtain

$$\mathcal{L}[P(s^2)] = \int_0^\infty e^{-zs^2} P(s^2) ds^2 = \varphi(z)^{-3/2} \quad (80)$$

and

$$\varphi(z) = \text{Det} |\mathbf{E} + \gamma z \mathbf{R}^{-1}| \quad (81)$$

where  $\gamma = 2b^2/(3N)$ . The Laplace transform of  $P(s^2)$ , i. e., the generating function of  $P(s^2)$ , provides the average  $\langle e^{-zs^2} \rangle$  and, therefore, the average of the powers of  $s^2$  can be computed by making use of the expansion in the form of a power series in  $z$ . Thus we have

$$\langle s^{2n} \rangle_0 = (-1)^n \frac{\partial^n}{\partial z^n} \varphi(z)^{-3/2} \Big|_{z \rightarrow 0}. \quad (82)$$

Since we have  $\text{Det} |\mathbf{R}| = N$  for any tree graph, (81) can be rewritten as

$$\varphi(z) = \frac{1}{N} \text{Det} |\mathbf{R} + \gamma z \mathbf{E}| = \frac{1}{N} \Phi(\mathbf{R}; -\gamma z). \quad (83)$$

Using (4) and (68), we have

$$\varphi(z) = \frac{1}{N} \text{Det} |\mathbf{A}_L + (\gamma z + 2) \mathbf{E}| = \frac{1}{N} \Phi(\mathbf{A}_L; -\gamma z - 2). \quad (84)$$

It was shown that the characteristic polynomial of the line graph gives the general equation for calculating the radius of gyration of a Gaussian chain with any type of branching [97].

According to Eichinger’s mathematical treatments [35], we showed the relation between eigenvalues of the matrix  $\mathbf{R}$  and the moments of the radius of gyration: ex-

pansion of (83) and comparing the coefficients of  $z$  give

$$\begin{aligned} \langle s^4 \rangle_0 - \langle s^2 \rangle_0^2 &= \frac{2b^4}{3N^2} \text{Tr}(\mathbf{R}^{-1})^2 = \frac{2b^4}{3N^2} \sum_{i=1}^{N-1} \frac{1}{\lambda_i^2} \\ &= \frac{2b^4}{3N^2} \sum_{i=1}^{N-1} \frac{1}{(\mu_i + 2)^2}, \end{aligned} \quad (85)$$

$$\begin{aligned} \langle s^6 \rangle_0 - 3\langle s^4 \rangle_0 \langle s^2 \rangle_0 + 2\langle s^2 \rangle_0^3 &= \frac{8b^6}{9N^3} \text{Tr}(\mathbf{R}^{-1})^3 \\ &= \frac{8b^6}{9N^3} \sum_{i=1}^{N-1} \frac{1}{\lambda_i^3} \\ &= \frac{8b^6}{9N^3} \sum_{i=1}^{N-1} \frac{1}{(\mu_i + 2)^3}. \end{aligned} \quad (86)$$

The above formulas also yield the relation between viscoelasticity and chain dimensions through the Rouse matrix.

The significance of the present graph-theoretical approach is to provide the general equations for the relaxation spectrum and the radius of gyration of any tree-like chain. In particular, it is noteworthy that the mathematical method has the potential to provide an algorithmic method of calculating high-order moments of the radius of gyration and the relaxation time for any tree-like chain. These values can hardly be calculated from the usual statistical methods because of a great difficulty in enumerating the distribution function.

The characteristic function (81) can be rewritten in terms of the  $N - 1$  eigenvalues of the Rouse matrix as [35]

$$\varphi(z) = \prod_{i=1}^{N-1} \left( 1 + \frac{\gamma}{\lambda_i} z \right). \quad (87)$$

Making use of the relation

$$\begin{aligned} \ln \varphi(z) &= \sum_{i=1}^{N-1} \ln \left( 1 + \frac{\gamma}{\lambda_i} z \right) \\ &= \sum_{i=1}^{N-1} \left[ \frac{\gamma}{\lambda_i} z - \frac{1}{2} \left( \frac{\gamma}{\lambda_i} \right)^2 z^2 + \frac{1}{3} \left( \frac{\gamma}{\lambda_i} \right)^3 z^3 - \dots \right], \end{aligned} \quad (88)$$

the sums of the reciprocal powers of the eigenvalues are easily determined from the following equation:

$$(n-1)! \sum_{i=1}^{N-1} \left( \frac{\gamma}{\lambda_i} \right)^n = (-1)^{n-1} \frac{\partial^n}{\partial z^n} \ln \varphi(z) \Big|_{z \rightarrow 0}. \quad (89)$$

Using (65), we have

$$\sum_{i=1}^{N-1} \tau_i^n = \frac{(-1)^{n-1}}{(n-1)!} \left( \frac{N\zeta_0}{4k_B T} \right)^n \frac{\partial^n}{\partial z^n} \ln \varphi(z) \Big|_{z \rightarrow 0}. \quad (90)$$

The coefficient of  $z$  in the characteristic function (81) can be related to a topological index such as the Wiener index of the molecular graph  $G$ , i.e., a total sum of the elements of its distance matrix, which is potentially useful in the correlation of molecular topology to thermodynamic properties for alkanes. Therefore, any high-order coefficient in (81) for a tree-like graph, or any coefficient of the characteristic polynomial for its line graph, has the potential to be a new topological index. From the coefficients, we can define the high-ordered Wiener indices as

$$W_k \equiv (k-1)! \left( \frac{2}{3} \right)^{k-1} N^k \sum_{i=1}^{N-1} \left( \frac{1}{\lambda_i} \right)^k. \quad (91)$$

One can easily see that the first index  $W_1$  becomes the original  $W$ . For example, the second and third Wiener indices,  $W_2$  and  $W_3$ , can be used for calculating the higher moments of radius of gyration and mechanical relaxation spectrum of branched chains. The second and third Wiener indices provide the fourth and sixth moments of radius of gyration.

$$\langle s^4 \rangle_0 = \left( \frac{b}{N} \right)^4 (W_2 + W_1^2), \quad (92)$$

$$\langle s^6 \rangle_0 = \left( \frac{b}{N} \right)^6 (W_3 + 3W_2 W_1 + W_1^3). \quad (93)$$

## Applications

In this section, the relationship between the characteristic polynomial and the mechanical relaxation spectra is examined. Then the zero-shear-rate viscosity  $\eta_0$  and the steady-state compliance  $J_e^0$  [38], a measure of the elastic energy stored under steady flow, are

$$\begin{aligned} \eta_0 &= \frac{ck_B T}{N} \sum_{i=1}^{N-1} \tau_i, \\ J_e^0 &= \frac{N}{ck_B T} \left( \frac{\sum_{i=1}^{N-1} \tau_i^2}{\left( \sum_{i=1}^{N-1} \tau_i \right)^2} \right) \end{aligned} \quad (94)$$

so that these basic rheological functions can be rewritten using the Wiener indices as

$$\begin{aligned} \eta_0 &= \frac{c\zeta_0}{6} \left( \frac{b}{N} \right)^2 W_1, \\ J_e^0 &= \frac{3N}{2ck_B T} \left( \frac{W_2}{W_1^2} \right). \end{aligned} \quad (95)$$

Thus, the rheological properties can be calculated from the distance matrix of the corresponding tree-graph. This is a very important result because we can realize the effects of the branching feature such as the branch length, branch number, and branch position on these rheological behaviors. In addition, the intrinsic dynamic moduli of storage  $[G']$  and loss  $[G'']$  can be obtained by using the following relations [129]:

$$[G'] = \sum_{i=1}^{N-1} \frac{\omega^2 \tau_i^2}{1 + \omega^2 \tau_i^2}, \quad [G''] = \sum_{i=1}^{N-1} \frac{\omega \tau_i}{1 + \omega^2 \tau_i^2}, \quad (96)$$

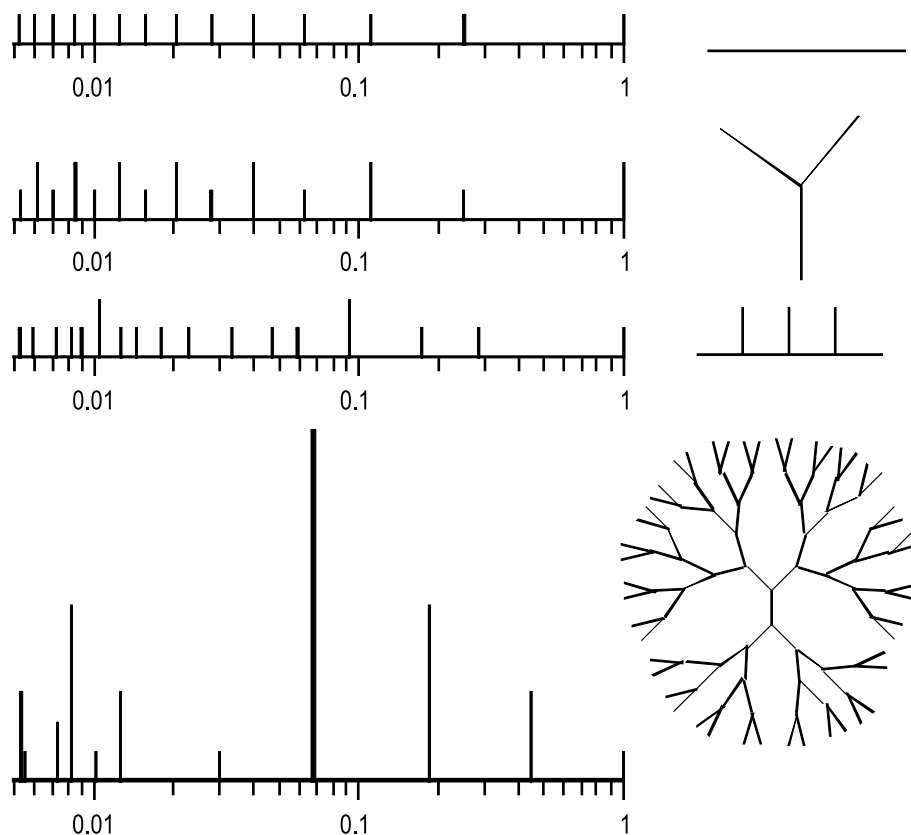
where the  $\omega$  is the frequency of strain oscillation.

It is possible to obtain the relaxation spectra through (71) from the eigenvalues of the adjacency matrix of its line graph. Calculations of relaxation spectra were exemplified for a linear, Y-shape star, comb, and dendrimer with a fixed vertex number of  $N = 94$  in Fig. 6. The corresponding frequency dependences of  $[G']$  and  $[G'']$  are plotted in Fig. 7 with the reduced variables. As seen in these figures, the low frequency slopes of  $[G']$  and  $[G'']$

are 2 and 1 on the logarithmic plot as expected for any linear polymer liquid but the higher frequency slope of  $[G']$  for the dendrimer was found to be larger than the 1/2 of the linear chain. It is evident that the comb and Y-star are almost intermediate between the linear and the dendrimer. In the flow region,  $\omega \rightarrow 0$ ,  $[G']$  and  $[G'']$  is going to  $\omega^2 \tau_R$  and  $\omega \tau_R$  as is obvious from (96) where  $\tau_R$  is the highest relaxation time or Rouse time. The Rouse time  $\tau_R$  can be estimated from the smallest eigenvalue of  $\mathbf{K}$  or  $\mathbf{R}$  or the second smallest one of  $\mathbf{L}^+$  or  $\mathbf{Z}$ . The order of  $\tau_R$  was in linear > comb > Y-star > dendrimer.

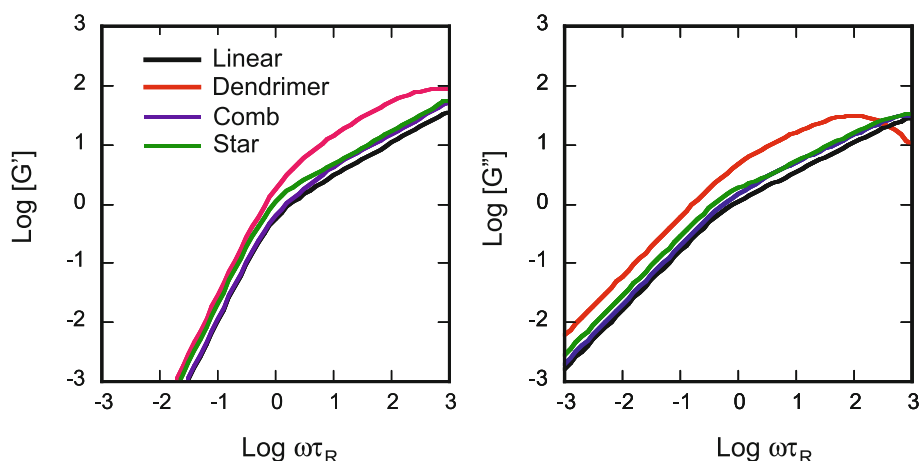
### Future Directions

The formulation of conformational statistics and dynamics due to the graph-theoretical expressions is based on the fact that the random-flight model and/or Gaussian chain model may be characterized by the so-called Markov nature such that the mean-square radius is proportional to the number of bonds in the chain in the unperturbed state without excluded volume. However, the Markov na-



Polymers, Non-linearity in, Figure 6

Relaxation spectra of linear, Y-star, comb, and dendrimer chains with a fixed vertex number of 94



### Polymers, Non-linearity in, Figure 7

Logarithmic plot of reduced intrinsic moduli  $G'$  and  $G''$  against  $\omega \tau_R$  for linear, comb, star, and dendrimer chains with  $N = 94$ , where  $\tau_R$  is the maximum relaxation time in Rouse dynamics

ture breaks down for real short chains and semiflexible or stiff chains [135]. This non-Markov nature arises from a sort of stiffness or static rigidity, as introduced by the constraints on the internal degrees of freedom such as fixed bond lengths, fixed bond angles, and hindered internal rotations. In other words, the Wiener index reflecting the radius of gyration based on the Markov nature can be considered to be an appropriate measure of molecular size and viscosity of flexible polymers rather than small molecules. For the study of such problems, the graph-theoretical studies for nonlinear non-Markov chains will play a central role in presenting the more useful devices and descriptors not only for semi-flexible chains and liquid crystalline polymers [98,135] but also for small molecules. The modification on the original bead-spring (Gaussian) chain or Markov chain model for nonlinear chains is to include the 'excluded volume effect' [129] because a substantial increase in the density of structural units close to branching point causes a stronger excluded-volume effects in comparison with other regions of the chain [25]. Regarding such theoretical difficulties inherent to branched structures, simulation work as well as the statistic-based methods such as renormalizing group theory [113] and self-avoiding random walk (SAW) procedure [77] will be a very powerful tool in the study of this type of branched polymers [48]. Furthermore, the combination with the graph-theoretical formulation will make it possible to find key parameters expressing the nature and degree of branching, which inevitably makes the mathematics more complicated.

The chain entanglements play an important role in the melt viscosity and rheological properties [54]. The extend of the theoretical framework for entanglements presented

by de Gennes [27,28] and Doi-Edwards [33] to branched polymers could be important for understanding the rheological properties of various branched polymers. Bonchev et al. [16] propose the topological descriptors of coarse-graining polymer graphs obtained by regarding the single edge as entanglement length. In particular the discrete mathematics such as graph theory and knot theory [121] seems to be important in the deeply understanding of the interplay between the entanglement and branching effects.

In addition, the effects of molecular weight distribution will be also the inevitable problems in industrial polymer products [4,92]. In random branching and hyperbranching processes the polydispersity index increases with the molecular weight [18,119]. The shape and width of the molar mass distribution curve remain extremely important also for branched polymers and markedly affect the melt properties of the polymers [86,87]. The concept of "forest" in graph theory will be appropriate to this problem.

Throughout this article, we have demonstrated that the graph-theoretical approach provides a topological insight into the nonlinearity in polymer architecture. This article deals with only the flexible polymers containing no loops nor rings are treated because of the lack of sufficient data on the statistics and dynamics of nonlinear chains with loops and rings. Thus, the problems of the dynamics and statistics of a tree-shaped molecule, which are very important for practical applications, were found to be completely reduced to the problem of the eigen-polynomial of the chain graph. This suggests that various ideas and concepts thus obtained from graph theory can be applied directly to the topological analysis for architecture in nonlinear polymers. For example, advantages in the formulation due



to the topological invariants are that the conformational statistics and dynamics of any branched polymer not only can be simply and universally evaluated, but also can be expressed as a function of topological parameters such as the position of branches and the length of main or side chains.

In conclusion, polymer chemistry would need a sense of the *mathematical chemistry* defined by Trinajstić–Gutman [128]: “*Mathematical chemistry is part of theoretical chemistry which is concerned with applications of mathematical methods to the chemical problems.*”. The topological descriptor widely used in polymer chemistry is the molecular weight or the degree of polymerization which seems to be an index with low discriminating power and appropriate only to linear chain molecules. The author believes that the topological sense resulting from discrete mathematics such as graph theory provides powerful devices for leading to the quantitative structure-property and structure-activity relationships in the respectable form.

## Bibliography

### Primary Literature

- Altenburg VK (1960) Zur Berechnung des Radius verzweigter Moleküle. *Kolloid-Z* 178:112–119
- Baesley JK (1953) The molecular structure of polyethylene. IV. Kinetic calculations of the effect of branching on molecular weight distribution. *J Am Chem Soc* 75:6123–6127
- Balakrishnan R (2004) The energy of a graph. *Linear Algebr Appl* 387:287–295
- Berger L, Meissner J (1992) Linear viscoelasticity, simple and planar melt extension of linear polybutadienes with bimodal molar mass distributions. *Rheol Acta* 31:63–74
- Berry GC, Orfino TA (1964) Branched polymers. III. Dimensions of chains with small excluded volume. *J Chem Phys* 40:1614–1621
- Bicerano J (1989) Molecular level calculations of the structures and properties of non-crystalline polymers. Dekker, New York
- Billmeyer FW (1953) The molecular structure of polyethylene. III. Determination of long chain branching. *J Am Chem Soc* 75:6118–6122
- Biswas P, Kant R, Blumen A (2000) Polymer dynamics and topology: Extension of stars and dendrimers in external fields. *Macromol Theor Simul* 9:56–67
- Bonchev D, Markel E, Dekmezian AH (2001) Topological analysis of long-chain branching patterns in polyolefins reciprocal distance matrix. *J Chem Inf Comput Sci* 41:1274–1285
- Bonchev D, Markel EJ, Dekmezian AH (2002) Long chain branch polymer chain dimensions: Application of topology to the Zimm–Stockmayer model. *Polymer* 43:203–222
- Bonchev D, Mekenyan O (1980) Topological approach to the calculation of the  $\pi$ -electron energy and energy gap of infinite conjugated polymers. *Z Naturforsch* 35a:739–747
- Bonchev D, Mekenyan O, Polansky OE (1981) Topological approach to the predicting of the electron energy characteristics of conjugated infinite polymers. II. PPP-calculations. *Z Naturforsch* 36a:643–646
- Bonchev D, Mekenyan O, Polansky OE (1981) Topological approach to the predicting of the electron energy characteristics of conjugated infinite polymers. III. The influence of some structural modifications of polymers. *Z Naturforsch* 36a:647–650
- Bonchev D, Mekenyan O, Protić G, Tranajstić N (1979) Application of topological indices to gas chromatographic data: Calculation of the retention indices of isomeric alkylbenzenes. *J Chromatogr* 176:149–156
- Bonchev D, Trinajstić N (1977) Information theory, distance matrix, and molecular branching. *J Chem Phys* 67:4517–4533
- Bonchev D, Dekmezian AH, Markel E, Faldi A (2003) Topology-rheology regression models for monodisperse linear and branched polyethylenes. *J Appl Polym Sci* 90:2648–265
- Bueche F (1954) The viscoelastic properties of plastics. *J Chem Phys* 22:603–609
- Burchard W (1972) Angular distribution of Rayleigh scattering from branched polycondensates. Amylopectin and Glycogen types. *Macromol* 5:604–610
- Burchard W (1999) Solution properties of branched macromolecules. *Adv Polym Sci* 143:113–194
- Casassa EF, Tagami Y (1969) An equilibrium theory for exclusion chromatography of branched and linear polymer chains. *Macromol* 2:14–26
- Cayley A (1874) On the mathematical theory of isomers. *Phylos Mag* 67:444–446
- Chompff AJ (1970) Normal modes of branched polymers. I. Simple ring and star-shaped molecules. *J Chem Phys* 53:1566–1576
- Chompff AJ (1970) Normal modes of branched polymers. II. Complex branched molecules and ring systems. *J Chem Phys* 53:1577–1584
- Cvetković D (2005) Signless Laplacian and line graph. *Bull Acad Serbe Sci Arts Cl Sci Math Natur Sci Math* 31:85–92
- Daout M, Cotton JP (1982) Star shaped polymers: A model for the conformation and its concentration dependence. *J Phys* 43:531–538
- Dawkins JV, Maddock JW, Coupe D (1970) Gel-permeation chromatography: Examination of universal calibration procedures for Polydimethylsiloxane in a poor solvent. *J Polym Sci A-2* 8:1803–1821
- De Gennes PG (1971) Reptation of a polymer chain in the presence of fixed obstacles. *J Chem Phys* 55:572–579
- De Gennes PG (1975) Reptation of stars. *J Phys* 36:1199–1203
- De Gennes PG (1979) Scaling concepts in polymer physics. Cornell Univ, New York
- Debye P (1946) The intrinsic viscosity of polymer solutions. *J Chem Phys* 14:636–639
- Dobson GR, Gordon M (1964) Configurational statistics of highly branched polymer systems. *J Chem Phys* 41:2389–2398
- Doi M (1974) Relaxation spectra of nonlinear polymers. *Polym J* 6:108–120
- Doi M, Edwards SF (1978) Dynamics of concentrated polymer systems. *J Chem Soc Faraday Trans* 2:1789–1832
- Ebrahimi KG, Takahashi M, Arai O, Masuda T (1995) Effects of molecular weight distribution on dynamic viscoelasticity and biaxial extensile flow behavior of polystyrene melts. *J Rheol* 39:1385–1397

35. Eichinger BE (1980) Configuration statistics of Gaussian molecules. *Macromol* 13:1–11
36. Eichinger BE (1976) Molecules as graphs. *J Polym Sci Symp* 54:127–134
37. Estrada E (1995) Edge adjacency relationships and a novel topological index related to molecular volume. *J Chem Inf Comput Sci* 35:31–33
38. Ferry JD (1980) *Viscoelastic properties of polymers*, 3rd edn. Wiley, New York
39. Fischer EW, Hahn K, Kugler J, Struth U, Born R (1984) An estimation of the number of tie molecules in semicrystalline polymers by means of neutron scattering. *J Polym Sci* 22:1491–1513
40. Fixman M (1962) Radius of gyration of polymer chains. *J Chem Phys* 36:306–318
41. Flory PJ (1937) Mechanism of vinyl polymerizations. *J Am Chem Soc* 59:236–241
42. Flory PJ (1953) *Principles of polymer chemistry*. Cornell, New York
43. Flory PJ (1969) *Statistical mechanics of chain molecules*. Hanser, New York
44. Flory PJ, Fox TG (1954) Treatment of intrinsic viscosities. *J Am Chem Soc* 73:1904–1908
45. Forsman WC (1968) Matrix methods for determining the dimensions of branched random-flight chains. *Macromol* 1:343–347
46. Forsman WC (1976) Graph theory and the statistics and dynamics of polymer chains. *J Chem Phys* 65:4111–4115
47. Freed KF (1972) Functional integrals and polymer statistics. *Adv Chem Phys* 22:1–128
48. Freire JJ (1999) Conformational properties of branched polymers: Theory and simulations. *Adv Polym Sci* 143:35–112
49. Fréchet JMJ (1994) Functional polymers and dendrimers: Reactivity, molecular architecture, and interfacial energy. *Science* 263:1710–1715
50. Fujimoto T, Narukawa H, Nagasawa M (1970) Viscoelastic properties of comb-shaped polystyrenes. *Macromol* 3:57–65
51. Gordon M (1979) From Riemann's metric to the graph metric, or applying Occam's razor to entanglements. *Polymer* 20:1349–1356
52. Gottlieb M, Bird BR (1976) A molecular dynamics calculation to confirm the incorrectness of the random-walk distribution for describing the Kramers freely jointed bead-rod chain. *J Chem Phys* 65:2467–2468
53. Graessley WW (1975) Statistical mechanics of random coil networks. *Macromol* 8:186–190
54. Graessley WW (1976) The entanglement concept in polymer rheology. *Adv Polym Sci* 16:1–179
55. Graessley WW, Masuda T, Roovers JEL, Hadjichristidis N (1976) Rheological properties of linear and branched polyisoprene. *Macromol* 9:127–141
56. Graovac A, Gutman I, John PE, Vidović, Vlah I (2001) On statistics of graph energy. *Z Naturforsch* 56a:307–311
57. Graovac A, Gutman I, Trinajstić N, Bonchev D (1972) Graph theory and molecular orbitals. Application of Sachs theorem. *Theoret Chim Acta* 26:67–78
58. Grest GS, Fetters LJ, Juang JS, Richter D (1996) Star polymers: Experiment, theory, and simulation. *Adv Chem Phys* 94: 67–163
59. Grosberg AY, Khokhlov AR (1994) *Statistical physics of macromolecules*. AIP, New York
60. Grubisic Z, Rempp P, Benoit H (1967) A universal calibration for gel permeation chromatography. *J Polym Sci B* 5:753–759
61. Guenza M, Perico A (1992) A reduced description of the local dynamics of star polymers. *Macromol* 25:5942–5949
62. Gutman I, Yeh Y-N, Lee S-L, Chen J-C (1994) Wiener numbers of dendrimers. *Commun Math Chem (MATCH)* 30:103–115
63. Gutman I, Yeh Y-N, Lee S-L, Luo Y-L (1993) Some recent results in the theory of the Wiener number. *Ind J Chem* 32A:651–661
64. Guttman CM, Hoffman JD, DiMarzio EA (1979) Monte Carlo calculation of SANS for various models of semicrystalline polyethylene. *Faraday Discuss Chem Soc* 68:297–309
65. Hadjichristidis N (1999) Synthesis of miktoarm star polymers. *J Polym Sci Polym Chem* 37:857–871
66. Hadjichristidis N, Pitsikalis M, Pispas S, Iatrou H (2001) Polymers with complex architecture by living anionic polymerization. *Chem Rev* 101:3747–3792
67. Ham JD (1957) Viscosity theory of branched and cross-linked polymers. *J Chem Phys* 26:625–633
68. Ham NS, Ruedenberg K (1958) Energy levels, atom populations, bond populations in the LCAO MO model and in the FE MO model. A quantitative analysis. *J Chem Phys* 29: 1199–1214
69. Harary F (1962) The determinant of the adjacency matrix of a graph. *SIAM Rev* 4:202–210
70. Harary F (1969) *Graph theory*. Addison-Wesley, Reading
71. Hosoya H (1971) Topological index, A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull Chem Soc Jpn* 44: 2332–2339
72. Hosoya H, Kawasaki K, Mizutani K (1972) Topological index and thermodynamic properties K Empirical rules on the boiling point of saturated hydrocarbons. *Bull Chem Soc Jpn* 45:3415–3421
73. Ivanciuc O, Balaban T, Balaban AT (1993) Reciprocal distance matrix, related local vertex invariants and topological indices. *J Math Chem* 12:309–318
74. Kim YH, Webster OW (1990) Water-soluble hyperbranched polyphenylene: A unimolecular micelle? *J Am Chem Soc* 112:4592–4593
75. Kirchhoff G (1847) Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird. *Ann Phys Chem* 72: 497–508
76. Kirkwood JG, Riseman J (1948) The intrinsic viscosities and diffusion constants of flexible macromolecules in solution. *J Chem Phys* 16:565–573
77. Klein J (1980) Self-avoiding walks constrained to strips, cylinders, and tubes. *J Stat Phys* 23:561–586
78. Klein J, Fletcher D, Fetters L (1983) Dynamics of entangled star-branched polymers. *Faraday Symp Chem Soc* 18: 159–171
79. Kloczkowski A (2002) Application of statistical mechanics to the analysis of various physical properties of elastomeric networks. *Polymer* 43:1503–1525
80. Kosmas MK, Gaunt DS, Whittington SG (1989) Dimensions of the branched of a uniform brush polymer. *J Phys A Math Gen* 22:5109–5116
81. Kramers HA (1946) The behavior of macromolecules in inhomogeneous flow. *J Chem Phys* 14:415–425

82. Kurata M, Abe M, Iwama M, Matsushima M (1972) Randomly branched polymers. I. Hydrodynamic properties. *Polym J* 3:729–738
83. Kurata M, Fukatsu M (1964) Unperturbed dimension and translational friction constant of branched polymers. *J Chem Phys* 41:2934–2944
84. Kurata M, Stockmayer WH (1963) Intrinsic viscosity and unperturbed dimensions of long chain molecules. *Adv Polym Sci* 3:196–312
85. Lovász L, Pelikán J (1973) On the eigenvalues of trees. *Period Math Hung* 3:175–182
86. Masuda T, Kitagawa K, Inoue T, Onogi S (1970) Rheological properties of anionic polystyrenes. II. Dynamic viscoelasticity of blends of narrow-distribution polystyrene. *Macromol* 3:116–125
87. Masuda T, Ohta Y, Onogi S (1981) Rheological properties of anionic polystyrenes. III. Characterization and rheological properties of four-branch polystyrenes. *Macromol* 4:763–768
88. McGraph JE (ed) (1981) Anionic polymerization: Kinetics, mechanism and synthesis, vol 166. ACS, Washington
89. Mekenyan O, Bonchev D, Trinajstić N (1980) Chemical graph theory: Modeling the thermodynamic properties of molecules. *Int J Quantum Chem* 18:369–380
90. Mekenyan O, Dimitrov S, Bonchev D (1963) Graph-theoretical approach to the calculation of physicochemical properties of polymers. *Eur Polym J* 19:1185–1193
91. Merris R (1990) The distance spectrum of a tree. *J Graph Theory* 14:365–369
92. Mills NJ (1969) The rheological properties and molecular weight distribution of polydimethylsiloxane. *Eur Polym J* 5:675–695
93. Milner ST (1991) Polymer brushes. *Science* 251:905–914
94. Mohar B (1991) Eigenvalues, diameter, and mean distance in graphs. *Graph Comb* 7:53–64
95. Mohar B (1993) A novel definition of the Wiener index for trees. *J Chem Inf Comput Sci* 33:153–154
96. Nitta K (1994) A topological approach to statistics and dynamics of chain molecules. *J Chem Phys* 101:4222–4228
97. Nitta K (1999) A graph-theoretical approach to statistics and dynamics of tree-like molecules. *J Math Chem* 25:133–143
98. Noda I, Hearst JE (1971) Polymer dynamics. V. The shear dependent properties of linear polymers including intrinsic viscosity, flow dichroism and birefringence, relaxation, and normal stresses. *J Chem Phys* 54:2342–2354
99. Orofino TA (1961) Branched polymers. II. Dimensions in non-interacting media. *Polymer* 2:305–314
100. Pearson DS, Rju VR (1982) Configurational and viscoelastic properties of branched polymers. *Macromol* 15:294–298
101. Peticolas WL (1963) Introduction to the molecular viscoelastic theory of polymers and its applications. *Rubber Chem Technol* 36:1422–1458
102. Platt JR (1947) Influence of neighbor bonds on additive bond properties in paraffins. *J Chem Phys* 15:419–420
103. Platt JR (1952) Prediction of isomeric differences in paraffin properties. *J Phys Chem* 56:328–336
104. Plavšić D, Nikolić S, Trinajstić N, Mihalić Z (1993) On the Harary index for the characterization of chemical graphs. *J Math Chem* 12:235–250
105. Polansky OE, Bonchev D (1987) The Wiener number of graphs. I. General theory and changes due to graph operations. *Commun Math Chem (MATCH)* 21:133–186
106. Polansky OE, Bonchev D (1990) Theory of the Wiener number of graphs. II. Transfer graphs and some of their metric properties. *Commun Math Chem (MATCH)* 25:3–40
107. Randić M (1975) On characterization of molecular branching. *J Am Chem Soc* 97:6609–6615
108. Roedal M (1953) The molecular structure of polyethylene. I. Chain branching in polyethylene during polymerization. *J Am Chem Soc* 75:6110–6112
109. Roovers J, Graessley WW (1981) Melt rheology of some model comb polystyrenes. *Macromol* 14:766–733
110. Rouse PE (1953) A theory of the linear viscoelastic properties of dilute solutions of coiling polymers. *J Chem Phys* 21:1272–1280
111. Rouvray DH (1975) The value of topological indices in chemistry. *Commun Math Chem (MATCH)* 1:125–134
112. Sack RA (1953) Mean square radius of randomly coiled molecular chain. *Nature* 171:310
113. Schäfer L (1999) Excluded volume effects in polymer solutions. Springer, Berlin
114. Sheridan PF, Adolf DB, Lyulin AV, Neelov I, Davies GR (2002) Computer simulations of hyperbranched polymers: The influence of the Wiener index on the intrinsic viscosity and radius of gyration. *J Chem Phys* 117:7802–7812
115. Sperati CA, Franta WA, Starkweather HW (1953) The molecular structure of polyethylene. V. The effect of chain branching and molecular weight on physical properties. *J Am Chem Soc* 75:6127–6133
116. Spialter L (1968) The atom connectivity matrix and its characteristic polynomial: A new computer-oriented chemical nomenclature. *J Am Chem Soc* 85:2012–2013
117. Stejskal J, Horská J, Kratochvíl P (1984) Graft copolymer statistics. *Macromol* 17:920–926
118. Stejskal J, Kratochvíl P, Jenkins AD (1987) Graft copolymer statistics. 2. Application to graft copolymers prepared from macromonomers. *Macromol* 20:181–185
119. Stockmayer WH (1943) Theory of molecular size distribution and gel formation in branched chain polymers. *J Chem Phys* 11:45–55
120. Stockmayer WH, Fixman M (1953) Dilute solutions of branched polymers. *Ann NY Acad Sci* 57:334–352
121. Summers DW (1987) Konts, macromolecules and chemical dynamics. In: King RB, Rouvray DH (eds) *Graph theory and topology in chemistry*. Elsevier, Amsterdam, pp 3–22
122. Sylvester JJ (1878) Chemistry and algebra. *Nature* 17:284
123. Sylvester JJ (1878) On an application of the new atomic theory to graphical representation of the invariants and covariants of binary quantics. *Amer J Math* 1:64–125
124. Szeleifer I, Carignano MA (1996) Tethered polymer layers. *Adv Chem Phys* 94:165–260
125. Tackx P, Tacx JCJF (1998) Chain architecture of LDPE as a function of molar mass using size exclusion chromatography and multi-angle laser light scattering (SEC-MALLS). *Polymer* 39:3109–3113
126. Tang H (1996) Rouse dynamics of block copolymers. *Macromol* 29:2633–2640
127. Tomalia DA, Baker H, Dewald J, Hall M, Kallos G, Martin S (1985) A new class of polymers: Starburst-dendritic macromolecules. *Polym J* 17:117–132

128. Trinajstić N, Gutman I (2002) Mathematical chemistry. *Croatia Chem Act* 75:329–356
129. Tschoegl NW (1964) Influence of hydrodynamic interaction on the viscoelastic behavior of dilute polymer solutions in good solvents. *J Chem Phys* 40:473–479
130. Ugi T, Marquarding D, Klusacek H, Gokel G, Gillespie P (1970) Chemie und logische Strukturen. *Angew Chem* 82:741–771
131. Van Dam ER, Haemers WH (2003) Which graphs are determined by their spectrum? *Linear Algebr Appl* 373:241–272
132. Widmann AH, Davies GR (1998) Simulation of the intrinsic viscosity of hydrobranched polymers with varying topology. 1. Dendritic polymers built by sequential addition. *Comput Theor Polym Sci* 8:191–199
133. Wiener H (1947) Correlation of heats of isomerization and differences in heats of vaporization of isomers among the paraffin hydrocarbons. *J Am Chem Soc* 69:2636–2638
134. Wiener H (1947) Structural determination of paraffin boiling points. *J Am Chem Soc* 69:17–20
135. Yamakawa H (1971) *Modeln theory of polymer solutions*. Harper and Row, New York
136. Yang Y (1998) Graph theory of viscoelastic and configurational properties of Gaussian chains. *Macromol Theory Simul* 7:521–549
137. Zimm BH (1953) Dynamics of polymer molecules in dilute solution: Viscoelasticity, flow birefringence and dielectric loss. *J Chem Phys* 24:269–278
138. Zimm BH, Kilb RW (1959) Dynamics of branched polymer molecules in dilute solution. *J Polym Sci* 37:19–42
139. Zimm BH, Stockmayer WH (1949) The dimensions and chain molecules containing branches and rings. *J Chem Phys* 17:1301–1314

### Books and Reviews

- Balaban AT (1976) *Chemical applications of graph theory*. Academic Press, London
- Beineka LW, Wilson RJ (2004) *Topics in algebraic graph theory*. Cambridge Univ, Cambridge
- Bonchev D, Rouvray DH (1991) *Chemical graph theory: Introduction and fundamentals*. Abasuc Press, Gordon and Breach Sci Publ, New York
- Doi M, Edwards SF (1986) *The theory of polymer dynamics*. Clarendon, Oxford
- Flory PJ (1969) *Statistical mechanics of chain molecules*. Wiley, New York
- King RB (1983) *Chemical applications of topology and graph theory*. Elsevier, Amsterdam
- Small PA (1975) Long-chain branching in polymers. *Adv Polym Sci* 18:1–64
- Trinajstić N (1992) *Chemical graph theory*, 2nd edn. CRC Press, Florida

## Popular Wavelet Families and Filters and Their Use

MING-JUN LAI

Department of Mathematics, The University of Georgia, Athens, USA

### Article Outline

Glossary  
 Introduction  
 Definition of Wavelets  
 Definition of Filters  
 Multi-Resolution Analysis  
 Wavelet Decomposition and Reconstruction  
 Refinable Functions  
 Compactly Supported Orthonormal Wavelets  
 Parameterization of Orthonormal Wavelets  
 Biorthogonal Wavelets  
 Prewavelets  
 Tight Wavelet Frames  
 Tight Wavelet Frames over Bounded Domain  
 q-Dilated Orthonormal Wavelets  
 Multiwavelets and Balanced Multiwavelets  
 Multivariate Orthonormal Wavelets  
 Biorthogonal Box Spline Wavelets  
 Multivariate Prewavelets  
 Multivariate Tight Wavelet Frames and Bi-Frames  
 Spherical Tight Wavelet Frames  
 Wavelets for Image Processing  
 Future Directions  
 Bibliography

### Glossary

**B-splines**  $N_d$  is the uniform B-spline of order  $d$  based on integer knot sequence. It is a function of piecewise polynomial of degree  $d - 1$  and smoothness  $d - 2$ . Trigonometric B-splines  $T_d$  will also be used.

**Box splines**  $B_{\ell,m,n}$  is a box spline of degree  $\ell + m + n - 2$  on three direction mesh.  $B_{k,\ell,m,n}$  is a box spline of degree  $k + \ell + m + n - 2$  on four direction mesh. They are bivariate piecewise polynomial functions of certain smoothness dependent on integers  $k, \ell, m, n$ .

**Filter** A filter is a sequence of real numbers. For example, a FIR filter is a finite sequence of real numbers. An IIR filter is a sequence of real numbers whose discrete Fourier transform is a rational function in  $z = e^{i\omega}$ .

**Filter process** A filter process is to convolute a digital signal with a filter, converting an input digital signal to an output digital signal. A subband coding scheme is a synthetic filter process which convert an input signal to several output signals.

**Image compression** A procedure to use less bytes of information to represent the same image (within tolerance). That is, for an image of size  $512 \times 512$  and standard integer gray level  $[0, 255]$ , the image needs a file of  $512 \times 512 \times 8$  bytes to store in a computer or to be sent over the internet. If one can use a file of some bytes less



than  $512 \times 512 \times 8$  to represent this image (storage or transmission), then the file is a compressed image.

**Image denoise** A procedure to remove noises from a noised image to make the image sharper and clearer.

**Image edge detection** A procedure to find features, skeleton or segmentation of images.

**$L_2$  spaces** A space of all square integrable functions.

**Mask** A mask is a finite sequence of real numbers. A mask polynomial is the discrete Fourier transform of a mask. Sometimes, a mask polynomial is also called the symbol of a mask.

**MRA** MRA stands for multi-resolution approximation (of a  $L_2$  space).

**Wavelet** A function or a group of functions which can generate a basis for Hilbert space by its translations and dilations is called wavelet. Many generalized versions of wavelets will be discussed including orthonormal wavelets, biorthogonal wavelets, pre-wavelets, wavelet frames, multi-wavelets, q-dilated wavelets, multivariate nonseparable wavelets.

## Introduction

Wavelets are one or a few functions whose integer translations and dilations can generate a basis for a Hilbert space. The concept of wavelets was introduced in the 1980's and has since been generalized and extended in many directions. The theory and applications have been continuously developed. One of its significant features is that it provides a systematical approach for designing various filters and filter banks for signal and image processing. Another feature is that wavelets leads to the theory of multi-resolution approximation (MRA). Wavelets and MRA have found many applications in most areas of science and technology, e. g., astronomy, electric engineering, fuzzy logic, geoscience, medical imaging, physics, and statistics. Wavelets have become an important subject in applied mathematics, approximation theory, numerical analysis and harmonic analysis.

In this article we present only the discrete wavelet transform, omitting the discussion of continuous wavelet transform. We shall describe various kinds of wavelets, outline their construction, and present some examples. To simplify the study of the regularity properties of wavelets, we restrict our discussion to primarily those wavelets whose construction is based on the well-known refinable functions, B-splines and box splines whose regularity is fully understood. Additionally, some wavelets without smoothness will be given as examples. Also, we shall describe the concepts of filters, filter banks, filtering processes, and their connection to wavelets. For the pur-

pose of applications, only finite impulse response filters and realizable infinite impulse response filters (some rational filters) can be implemented. Thus, we shall restrict our attention to compactly supported wavelet functions with the exception of those associated to stable rational filters. All globally supported wavelet functions such as Meyer's wavelet, Battle-Lemarie's wavelets, Shannon's wavelets, and Freedman's spherical wavelets will not be mentioned further in this article. Also we only discuss real-valued wavelets although there are complex-valued wavelets available in the literature.

For applications of wavelets, we will consider three applications in image processing. Other applications in signal processing, in solution of integral equations and partial differential equations, in subdivision algorithms for curves and surfaces, and in statistics, have to be omitted due to the space and time limitation.

## Definition of Wavelets

Let us start with a well-known Hilbert space,  $L_2(\mathbb{R})$ , the space of all square integrable functions on  $\mathbb{R}$ . That is,

$$L_2(\mathbb{R}) = \left\{ f : \int_{-\infty}^{\infty} |f(x)|^2 dx < +\infty \right\}.$$

A function  $\psi \in L_2(\mathbb{R})$  is a *wavelet* if the integer translations and dilations

$$\psi_{jk}(x) := 2^{j/2} \psi(2^j x - k), \quad j, k \in \mathbb{Z} \quad (1)$$

of  $\psi$  form a basis for  $L_2(\mathbb{R})$ . That is, these  $\psi_{j,k}$  are linearly independent, and any function  $f \in L_2(\mathbb{R})$  can be represented as a linear combination of  $\psi_{jk}$ .

It is clear that such a function is useful for computation since one only needs to store one function  $\psi$  in computer to represent any function in  $L_2(\mathbb{R})$ . In particular, a compactly supported wavelet  $\psi$  is more appropriate for the evaluation of any  $f$  represented in terms of these  $\psi_{jk}$ .

There are many kinds of wavelet functions. When these  $\psi_{jk}$  are orthonormal, i. e.,

$$\int_{-\infty}^{\infty} \psi_{j,k}(x) \psi_{m,n}(x) dx = \begin{cases} 1, & \text{if } j = m, \quad k = n \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$\psi$  is called *orthonormal wavelet*.

Let  $W_j = \text{span}\{\psi_{jk}, k \in \mathbb{Z}\}$  for each  $j$ . When  $W_j$  is orthogonal to  $W_m$  for  $m \neq j$  while the integer translates  $\psi_{jk}, k \in \mathbb{Z}$  may not be orthonormal to each other,  $\psi$  is called *prewavelet* or *semi-orthonormal wavelet*.



Since  $\psi_{jk}$ ,  $j, k \in \mathbb{Z}$  form a basis, for each  $f \in L_2(\mathbb{R})$

$$f = \sum_{j,k \in \mathbb{Z}} c_{jk} \psi_{jk}$$

for some coefficients  $c_{jk}$ . Suppose that there are two positive constants  $A$  and  $B$  such that

$$A \sum_{j,k \in \mathbb{Z}} |c_{jk}|^2 \leq \left\| \sum_{j,k \in \mathbb{Z}} c_{jk} \psi_{jk} \right\|_2^2 \leq B \sum_{j,k \in \mathbb{Z}} |c_{jk}|^2$$

for all coefficients  $c_{jk}$ . Then  $\psi$  is called *Riesz wavelet*.

Suppose that there exists another function  $\tilde{\psi} \in L_2(\mathbb{R})$  associated with  $\psi$  such that

$$\int_{-\infty}^{\infty} \psi_{j,k}(x) \tilde{\psi}_{m,n}(x) dx = \begin{cases} 1, & \text{if } j = m \text{ and } k = n \\ 0, & \text{otherwise} \end{cases}$$

for  $j, k, m, n \in \mathbb{Z}$ , where  $\tilde{\psi}_{jk}(x) = 2^{j/2} \tilde{\psi}(2^j x - k)$ . Suppose that both  $\psi$  and  $\tilde{\psi}$  are Riesz wavelets. Then  $\psi$  is called *biorthogonal wavelet* and  $\tilde{\psi}$  is a dual of  $\psi$ . In this case,

$$f = \sum_{j,k \in \mathbb{Z}} \langle f, \tilde{\psi}_{jk} \rangle \psi_{jk} = \sum_{j,k \in \mathbb{Z}} \langle f, \psi_{jk} \rangle \tilde{\psi}_{jk}$$

for all  $f \in L_2(\mathbb{R})$ , where

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x) g(x) dx$$

is the standard inner product in  $L_2(\mathbb{R})$ . Clearly, when  $\tilde{\psi} = \psi$ ,  $\psi$  is an orthonormal wavelet.

Furthermore, let us now assume that  $\psi_{jk}$ ,  $j, k \in \mathbb{Z}$ , form a redundant basis for  $L_2(\mathbb{R})$  in the sense that any function  $f \in L_2(\mathbb{R})$  can be expressed by these  $\psi_{jk}$  in the following sense

$$f = \sum_{j,k \in \mathbb{Z}} c_{jk} \psi_{jk},$$

where these  $\psi_{jk}$  may not be linearly independent and  $c_{jk}$  not unique. We say that these  $\psi_{jk}$  form a *tight wavelet frame* if

$$\sum_{j,k \in \mathbb{Z}} |\langle f, \psi_{jk} \rangle|^2 = \|f\|^2$$

for all  $f \in L_2(\mathbb{R})$ , where  $\|f\|^2 := \langle f, f \rangle$ .

Next, let us consider a Sobolev space  $H^k(\mathbb{R})$  consisting of all functions  $f$  whose derivatives up to  $k > 0$  are in  $L_2(\mathbb{R})$ . That is,

$$H^k(\mathbb{R}) = \{f: f^{(r)} \in L_2(\mathbb{R}), r = 0, \dots, k\},$$

where  $f^{(r)}$  denotes the  $r$ th derivative of  $f$ . Certainly, if we replace  $L_2(\mathbb{R})$  by a Sobolev space  $H^k(\mathbb{R})$  above, we will get corresponding Sobolev wavelets. In particular, if we let  $\psi$  be a function in  $H^k(\mathbb{R})$  and  $\psi_{jk}$  defined as in (1), and if

$$\langle \psi_{jk}, \psi_{mn} \rangle_k = \begin{cases} 1, & \text{if } j = m \text{ and } k = n \\ 0, & \text{otherwise} \end{cases},$$

then  $\psi$  is called *orthonormal Sobolev wavelet*. Here  $\langle f, g \rangle_k$  stands for the inner product in  $H^k(\mathbb{R})$  defined by

$$\langle f, g \rangle_k = \sum_{r=0}^k \langle f^{(r)}, g^{(r)} \rangle$$

for all  $f, g \in H^k(\mathbb{R})$ . Similarly we can define biorthogonal wavelets in Sobolev spaces (cf. [3,4,41]).

Furthermore, we can generalize the concepts of wavelets in several directions. If we consider the  $d$ -dimensional Euclidean space  $\mathbb{R}^d$  with  $d > 1$ , then all of the above concepts for various wavelets can be defined in  $L_2(\mathbb{R}^d)$  or in  $H^k(\mathbb{R}^d)$  accordingly. These concepts will be explained in detail in later sections.

Another direction of generalization is to consider a group of wavelet functions. Let  $\Psi$  be a function vector of finite length  $r > 1$ . That is,

$$\Psi = [\psi_1, \dots, \psi_r]^T.$$

For example,  $\Psi$  contains a symmetric function  $\psi_1$ , anti-symmetric function  $\psi_2$ , a sufficiently smooth function  $\psi_3$  and a discontinuous function  $\psi_4$ . In fact, a wavelet vector  $\Psi$  consisting of functions of various shapes and properties is very useful to represent any function  $f \in L_2(\mathbb{R})$ .

A new direction of generalization is to consider a dilation factor  $s > 2$ . That is, we define

$$\psi_{s,j,k}(x) = s^{j/2} \psi(s^j x - k), \quad j, k \in \mathbb{Z}.$$

In this case, one wavelet function is usually not enough. For example, when  $s = 3$  we will need 2 wavelet functions.

The above two generalizations lead to the following concept of multi-wavelets. Let  $\Psi_{s,j,k}$  be the  $k$ th translation and  $j$ th dilation of  $\Psi$  with integer dilation factor  $s \geq 2$ . That is,

$$\Psi_{s,j,k} = [s^{j/2} \psi_1(s^j x - k), \dots, s^{j/2} \psi_r(s^j x - k)]^T.$$

Suppose that any function  $f \in L_2(\mathbb{R})$  can be expressed by using the entries of  $\Psi_{s,j,k}$ , i.e., there exist coefficient vectors  $c_{j,k}$  of length  $r$  such that

$$f = \sum_{j,k \in \mathbb{Z}} c_{j,k}^T \Psi_{s,j,k}.$$

Suppose that we have

$$\int_{-\infty}^{\infty} \Psi_{s,j,k}(x) \Psi_{s,m,n}(x)^T dx = I_r \delta_{jm} \delta_{kn}$$

with  $I_r$  being the identity matrix of size  $r \times r$  and  $\delta_{jm}$  being the Kronecker delta, i. e.,  $\delta_{jm} = 1$  if  $j = m$  and  $\delta_{jm} = 0$  otherwise. Then  $\Psi$  is called an *orthonormal multi-wavelet*. Other concepts of wavelets can be generalized in this multi-wavelet setting as well. We leave the discussion to later sections.

Yet, another direction of generalization is to consider wavelet functions over a bounded domain. Let  $\Omega \subset \mathbb{R}^d$  be a bounded domain, e. g.,  $\Omega = [0, 1]$  when  $d = 1$ . One would like to have locally supported wavelet functions over  $\Omega$ .

### Definition of Filters

We begin with the following basic concepts in signal processing. A continuous-time signal, or sometimes called an analog signal,  $u(t)$ , is a piecewise continuous function of the time variable  $t$ , where  $t$  ranges from  $-\infty$  to  $\infty$ . It is called a band-limited signal if

$$u(t) = \int_{-\omega_0}^{\omega_0} \sigma(\omega) e^{i\omega t} d\omega \quad (3)$$

for a function  $\sigma(\omega)$  in  $L^1(-\omega_0, \omega_0)$ , where  $\omega_0$  is a positive number. Certainly we shall consider only signals  $u(t)$  of finite energy, i. e.,

$$\int_{-\infty}^{\infty} |u(t)|^2 dt < \infty.$$

In mathematical language, we shall restrict ourselves to functions in  $L_2(\mathbb{R})$ . If a band-limited analog signal  $u(t)$  happens to be in  $L_1(\mathbb{R})$ , the space of all absolutely integrable functions over  $\mathbb{R}$ , then the Fourier transform of  $u(t)$ , defined by

$$\hat{u}(\omega) = \int_{-\infty}^{\infty} u(t) e^{-i\omega t} dt$$

is given by

$$\hat{u}(\omega) = \begin{cases} 2\pi\sigma(\omega), & \text{for } -\omega_0 \leq \omega \leq \omega_0 \\ 0, & \text{otherwise} \end{cases}.$$

The Fourier transform takes  $u(t)$  defined on the *time domain* to  $\hat{u}(\omega)$  defined on the *frequency domain*. Here,

$\omega$  will be reserved for the frequency variable and  $i$  for the pure imaginary number  $\sqrt{-1}$ . The length of the smallest subinterval of  $(-\omega_0, \omega_0)$  outside which  $\sigma(\omega)$  vanishes identically is called the *bandwidth* of  $u(t)$ . Hence, the bandwidth of  $u(t)$  as defined in (3) does not exceed  $2\omega_0$ . That is, the signal  $u(t)$  contains no frequency higher than  $\frac{\omega_0}{\pi}$  cycles per second.

Any analog signal  $u(t)$  can be converted to a discrete-time signal  $u_n$  ( $n = \dots, -1, 0, 1, \dots$ ) by first sampling  $u(t)$  periodically with sampling time  $t_0 > 0$  and then quantizing by rounding off the values of  $u(nt_0)$ .

It is important to note, however, that if the sampling time  $t_0$  is not chosen small enough, then the analog signal would not be well represented. For a band-limited analog signal, the following famous result gives us a guide-line for choosing  $t_0$ .

**Theorem 1 (Shannon's Sampling Theorem)** *Let  $u(t)$  be a band-limited analog signal in  $L_2(-\infty, \infty)$  with bandwidth  $2\omega_0$ , and let*

$$0 < t_0 \leq \frac{\pi}{\omega_0}.$$

*Then  $u(t)$ ,  $-\infty < t < \infty$ , can be recovered from its values  $u(nt_0)$ ,  $n = \dots, -1, 0, 1, \dots$ , by using the formula*

$$u(t) = \sum_{n=-\infty}^{\infty} u(nt_0) \frac{\sin \pi(t/t_0 - n)}{\pi(t/t_0 - n)},$$

*where the convergence is uniform in  $t \in (-\infty, \infty)$ .*

Of course, most analog signals are not band-limited. However, we have the following result:

**Theorem 2 (Plancherel's Theorem)** *There is a linear isometry  $\Psi$  of  $L_2(\mathbb{R})$  onto  $L_2(\mathbb{R})$  which is uniquely determined by the requirement that*

$$\Psi f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$$

*for every function  $f \in L_2(\mathbb{R}) \cap L_1(\mathbb{R})$  such that for every  $f, g \in L_2(\mathbb{R})$ ,*

$$\int_{-\infty}^{\infty} f(t) g(t) dt = \int_{-\infty}^{\infty} \Psi f(\omega) \Psi g(\omega) d\omega.$$

By the above theorem, if  $u(t)$  is in  $L_2(\mathbb{R})$  and has a Fourier transform (e. g., its frequency can be seen or measured), then its Fourier transform  $\hat{u}(\omega)$  is in  $L_2(\mathbb{R})$ . If  $u(t) \in L_2(\mathbb{R}) \cap L^1(\mathbb{R})$ , then  $\hat{u}(\omega)$  is a continuous function

on  $(-\infty, \infty)$ . Thus,  $\hat{u}(\omega)$  decays to 0 as  $\omega$  tends to  $\pm\infty$ , and  $\hat{u}(\omega)$  can be approximated by the truncated functions

$$\sigma_b(\omega) = \begin{cases} \frac{1}{2\pi} \hat{u}(\omega) & \text{for } -b < \omega < b \\ 0, & \text{otherwise,} \end{cases}$$

and this in turn implies that  $u(t)$  can be approximated by the band-limited signals

$$u_b(t) = \int_{-b}^b \sigma_b(\omega) e^{j\omega t} d\omega,$$

where  $b \in (0, \infty)$  is usually determined by some practical criteria.

Next we introduce the  $z$ -transform of a digital signal  $\{u_n: n \in \mathbb{Z}\}$  by

$$Z(\{u_n\}) = \sum_{n \in \mathbb{Z}} u_n z^{-n}$$

for  $z = re^{i\omega}$  for some  $0 < r < \infty$ , where  $u_n = u(nt_0)$  for a band-limited signal with bandwidth  $2\pi/t_0$ . Sometimes, it is called the discrete Fourier transform when  $r = 1$ .

The connection between an analog signal  $u(t)$  in the frequency domain and its digital signal  $\{u(nt_0)\}$  in the frequency domain is the following

**Theorem 3** Suppose that  $u(t) \in L^1(-\infty, \infty)$  and its Fourier transform  $\hat{u}$  of  $u$  is in  $L^1(-\infty, \infty)$ . Then

$$U^*(\omega) = \sum_n u(nt_0) e^{j\omega n t_0} = \frac{1}{t_0} \sum_n \hat{u}\left(\frac{2n\pi}{t_0} - \omega\right)$$

where  $t_0 > 0$  is a sampling time.

This can be verified by using the well-known Poisson's summation formula (cf. [8]).

With the above knowledge of analog and digital signals, let us explain filters and filtering process. Although a digital filter is a (linear or nonlinear) transformation that takes any digital signal  $\{u_n\}$ , called an input signal, to a digital signal  $\{v_n\}$ , called the corresponding output signal, we are only interest in those filters which are stable, time-invariant, and linear filters. Such a filter is uniquely determined by a sequence of complex numbers

$$\dots, h_{-2}, h_{-1}, h_0, h_1, h_2, \dots$$

with  $\sum_{n \in \mathbb{Z}} |h_n| < \infty$  and by the *filtering process* defined by convolution:

$$\{v_n\} = \{h_n\} * \{u_n\} \text{ with } v_n = \sum_{j=-\infty}^{\infty} h_j u_{n-j}. \quad (4)$$

By using the  $z$ -transform to both sides of the equation above, a digital filter is associated with a transfer function  $H(z)$  such that

$$Z(\{v_n\}) = H(z)Z(\{u_n\})$$

with  $H(z) = \sum_{n \in \mathbb{Z}} h_n z^{-n}$ .

For example, a filter such that for any input signal  $\{u_n\}$ , an output signal  $\{v_n\}$  is obtained as defined by

$$v_n = \frac{1}{k} \sum_{i=1}^k u_{n-i}$$

is an example of moving average filter. As another example, a filter which takes input signal  $\{u_n\}$  and output a digital signal  $\{v_n\}$  with  $v_n = u_{-n}$  for all  $n$  is called a mirror filter. And as another example, a transfer function  $H(z)$  satisfying  $|H(z)|^2 + |H(-z)|^2 = 1$  is called *conjugate filter*.

If  $\{h_n\}$  is a finite sequence (i.e. only finitely many  $h_n \neq 0$ ), it is called a *finite impulse response* (FIR) digital filter. If infinitely many  $h_n$  are nonzero, the filter is called an *infinite impulse response* (IIR) digital filter.

An FIR digital filter is easy to implement. For example, if  $h_j, j = 0, \dots, N$  are only nonzero, then

$$v_n = \sum_{j=0}^N h_j u_{n-j} = h_0 u_n + \dots + h_N u_{n-N}, \quad \forall n \in \mathbb{Z}.$$

Of course the operations described above can be considered as a weighted average with weights  $h_0, \dots, h_N$ , an FIR filter is also called a *moving average* (MA) digital filter.

On the other hand, an IIR digital filter cannot be implemented in the same manner, simply because it is not possible to implement infinitely many scalar multiplications except for those filters whose  $z$ -transform is a rational function in  $z$ . That is, we will see that a rational function  $H(z)$  indeed provides a realizable IIR digital filter. Suppose that

$$H(z) = \frac{a_0 + a_1 z^{-1} + \dots + a_M z^{-M}}{1 - b_1 z^{-1} - \dots - b_N z^{-N}},$$

where  $a_0, \dots, a_M, b_1, \dots, b_N$  are complex numbers and  $M, N$  are nonnegative integers. In terms of  $z$ -transform, we may write

$$\begin{aligned} & \left(1 - \sum_{n=1}^N b_n z^{-n}\right) \left(\sum_{n=0}^{\infty} v_n z^{-n}\right) \\ &= \left(\sum_{n=0}^M a_n z^{-n}\right) \left(\sum_{n=0}^{\infty} u_n z^{-n}\right). \end{aligned}$$

Comparing with the coefficients from both sides of the above equation, we obtain

$$v_n - \sum_{j=1}^N b_j v_{n-j} = \sum_{j=0}^M a_j u_{n-j}$$

or

$$v_n = \sum_{j=0}^M a_j u_{n-j} + \sum_{j=1}^N b_j v_{n-j}.$$

Observe that the filtered outputs  $v_{n-j}$ ,  $j = 1, \dots, N$ , are used again to give the output  $v_n$ . This shows that a rational transfer function provides a realizable digital filter.

However, an important question is that this filter must be stable, i. e., it transforms any bounded input digital signal to a bounded output digital signal. The following result provides a criterion (cf. [8]).

**Theorem 4 (Stability Criterion for IIR Causal Digital Filters)** *An IIR digital filter with transfer function  $H(z)$  is stable if and only if all the poles of the rational function  $H(z)$  lie in the open unit disk  $|z| < 1$ .*

Since any FIR filter is stable, more FIR filters than realizable IIR filters have been designed in practice.

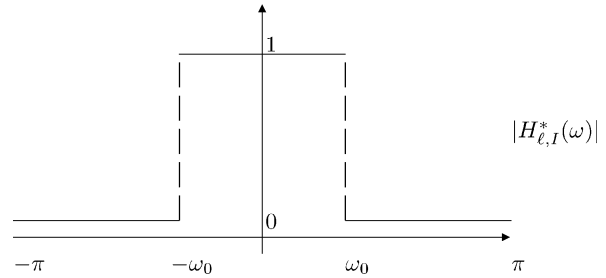
The most important role that the transfer function  $H(z)$  of a digital filter plays is to decide what frequencies to pass and what frequencies to stop. In practice, a whole range of frequencies must be filtered. For instance, if we wish to stop all frequencies  $\omega$  in the range  $a < \omega < b$  and pass those frequencies  $\omega$  in the range  $c < \omega < d$ , (where the intervals  $(a, b)$  and  $(c, d)$  lie in  $(0, \pi)$  and do not overlap), then we require

$$H^*(\omega) = \begin{cases} 0, & \text{for } a < \omega < b \\ 1, & \text{for } c < \omega < d. \end{cases}$$

The intervals  $(a, b)$  and  $(c, d)$  are called the *stopband* and *passband* of the digital filter, respectively. See Figs. 1 and 2

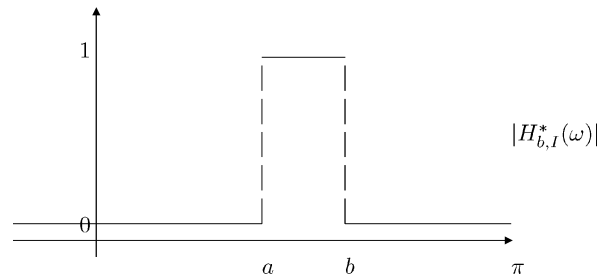
Unfortunately, any ideal amplitude filter characteristic  $|H_I^*(\omega)|$  which has a stopband (consisting of at least one interval) can not have a causal representation nor be an FIR filter. Thus we have to use FIR or realizable IIR filters to approximate the ideal filters  $H_I^*(\omega)$ . There are many ways to do such approximation. Wavelets are one of the modern approaches.

Subband coding is a good example of digital filtering application in signal processing. In Fig. 3, a subband coding scheme is shown, where  $[H]$  denotes the filter  $H$  which is convoluted with input signal  $\{s_n\}$ .  $[G]$ ,  $[H^\circ]$ , and  $[G^\circ]$  denote filters similarly.  $[2\downarrow]$  denotes a downsampling by



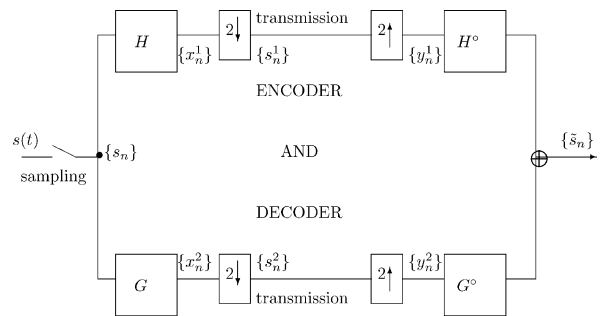
Popular Wavelet Families and Filters and Their Use, Figure 1

Ideal low-pass filter



Popular Wavelet Families and Filters and Their Use, Figure 2

Ideal band-pass filter



Popular Wavelet Families and Filters and Their Use, Figure 3

A subband coding schematic

2 of an input signal. That is, retain every even indexed sample values and delete every odd indexed sample values of an input digital signal. And  $[2\uparrow]$  denotes a upsampling by 2 of an input signal. That is, insert zero between every sample values of input signal.  $\oplus$  stands for a simple addition of two input digital signals.

A typical subband coding involves filtering a broad-band signal into two frequency bands (high-pass and lower-pass bands) so that the resulting subbands can be independently encoded for transmission. The encoders can be optimized to the statistics of the input signal, and made to give preferential coding weight to signal components in the parts of the spectrum that are perceptually most signif-

icant. The received signals are decoded and the subband components recombined to give an output signal which, for a given bit rate, is subjectively better than an equivalent broad-band waveform encoded signal. Ideally,  $\tilde{s}_n = s_{n+k}$  for all  $n$  by using four filters  $H, G, H^\circ$ , and  $G^\circ$ , where  $k$  is a certain time delay. This is called a perfect reconstruction. For application purposes, we say those four filters realize a perfect reconstruction if  $S(z) = \tilde{S}(z)$ , where  $S(z)$  and  $\tilde{S}(z)$  denote the  $z$ -transform of  $\{s_n\}$  and  $\{\tilde{s}_n\}$ .

Let  $X^1(z), X^2(z), Y^1(z), Y^2(z), S^1(z)$ , and  $S^2(z)$  denote the  $z$ -transform of digital signals  $\{x_n^1\}, \{x_n^2\}, \{y_n^1\}, \{y_n^2\}, \{s_n^1\}$ , and  $\{s_n^2\}$ , respectively. Write

$$H(z) = \sum_k h_k z^k \quad \text{and} \quad H^\circ(z) = \sum_k h_k^\circ z^k$$

and

$$G(z) = \sum_k g_k z^k \quad \text{and} \quad G^\circ(z) = \sum_k g_k^\circ z^k.$$

Then  $X^1(z) = H(z)S(z)$  and  $X^2(z) = G(z)S(z)$ . After down-sampling by 2,

$$s_n^1 = \sum_k h_{k-2n} s_k \quad \text{and} \quad s_n^2 = \sum_k g_{k-2n} s_k.$$

In terms of  $z$ -transform,

$$S^1(z^2) = \frac{1}{2}(X^1(z) + X^1(-z)) \quad \text{and} \\ S^2(z^2) = \frac{1}{2}(X^2(z) + X^2(-z)).$$

The received signal, after decoding, is up-sampled by 2 by inserting a zero valued sample between each received sample. That is,

$$y_n^1 = \begin{cases} s_{n/2}^1, & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd} \end{cases}$$

and

$$y_n^2 = \begin{cases} s_{n/2}^2, & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd} \end{cases}.$$

Or,  $Y^1(z) = S^1(z^2)$  and  $Y^2(z) = S^2(z^2)$ . Then the two subbands are filtered by  $H^\circ$  and  $G^\circ$ , respectively and are added into an output signal  $\{\tilde{s}_n\}$ . That is,

$$\tilde{S}(z) = H^\circ(z)Y^1(z) + G^\circ(z)Y^2(z).$$

Equivalently,

$$\tilde{s}_n = \sum_k h_{2k-n}^\circ s_k^1 + g_{2k-n}^\circ s_k^2, \forall n.$$

It can be easily seen that

$$\tilde{S}(z) = \frac{1}{2}[H^\circ(z)H(z) + G^\circ(z)G(z)]S(z) \\ + \frac{1}{2}[H^\circ(z)H(-z) + G^\circ(z)G(-z)]S(-z).$$

Thus, in order to achieve the perfect reconstruction, we need to choose  $H, G, H^\circ$ , and  $G^\circ$  such that  $\tilde{S}(z) = S(z)$ . In terms of those four filters, we need to have

$$\begin{cases} H^\circ(z)H(z) + G^\circ(z)G(z) & = 2 \\ H^\circ(z)H(-z) + G^\circ(z)G(-z) & = 0. \end{cases} \quad (5)$$

Hence, the above set of equations in (5) is called the perfect reconstruction condition.

*Example 1* If  $G(z) = H(-z)$ ,  $H^\circ(z) = H(z)$ , and  $G^\circ(z) = -H(-z)$ , then the second equation of the above system can be easily seen to satisfy. The first equation becomes

$$(H(z))^2 + (H(-z))^2 = 2.$$

Such a group of four filters is called a *Quadrature Mirror Filter Bank*. It is not easy to obtain such a bank of four filters. In practice, one designs filters so that

$$|(H(z))^2 + (H(-z))^2 - 2| \approx 0.$$

*Example 2* Consider  $G(z) = \overline{zH(-z)}$ . Choosing  $H^\circ(z) = \overline{H(z)}$  and  $G^\circ(z) = \overline{G(z)}$  with  $z = e^{-i\omega}$ , we have those four filters satisfied the second equation of the system (5) above. The first equation becomes

$$|H(z)|^2 + |H(-z)|^2 = 2.$$

Such a group of four filters is called a *Conjugate Quadrature Filter Bank*. In later sections we shall use wavelets to give many examples of such a bank of filters.

In the above, we have discussed a simple subband splitting. Certainly, we can further split each of bandpass signals before the encoder and transmission in the above diagram. If we split only low-pass signal in the above diagram and keep the high-pass signal unsplit, we get a nonuniform subband split. In this way, we shall obtain a wavelet decomposition of a signal if we use the filters associated with a wavelet. If we split both low- and high-pass signal further, we obtain a wavelet packet decomposition (cf. [91]).

It is easy to see that the concepts of filters, filtering process, subband coding scheme, and perfect reconstruction



have counterparts for multi-dimensional signals, e. g., 2D images (cf., e. g., [90]).

### Multi-Resolution Analysis

A simple example of orthonormal wavelet is *Haar wavelet*  $\psi$  defined by

$$\psi(x) = \begin{cases} 1, & \text{if } x \in [0, 1/2) \\ -1, & \text{if } x \in [1/2, 1) \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to check that  $\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$  are orthonormal and are dense in  $L_2(\mathbb{R})$ .

To construct a wavelet with more smoothness than that of the Haar wavelet we need a concept called multi-resolution approximation of  $L_2(\mathbb{R})$ . It is natural to let  $W_j = \text{span}\{\psi_{j,k}, k \in \mathbb{Z}\}$  and  $V_j = \bigcup_{-\infty < i \leq j} W_i$  for all  $j \in \mathbb{Z}$ . We can easily see that  $V_j \subset V_{j+1}$ . In fact, we have  $V_{j+1} = \{f(2x - k), \forall f \in V_j, k \in \mathbb{Z}\}$ . Suppose that  $V_0$  is spanned by translates of one function  $\phi$ , i. e.,

$$V_0 = \text{span} \left\{ \sum_{k \in \mathbb{Z}} c_k \phi(x - k), \quad c_k \in \mathbb{R} \right\}.$$

Then since  $V_0 \subset V_1$ , we have

$$\phi(x) = \sum_{k \in \mathbb{Z}} p_k \phi(2x - k) \quad (6)$$

which is called *dilation equation* and  $\phi$  is called *refinable function* or *scaling function*. In particular, if  $\phi$  is also orthonormal in the sense that

$$\int_{-\infty}^{\infty} \phi(x) \phi(x - k) dx = \delta_{0k}, \quad \forall k \in \mathbb{Z}, \quad (7)$$

then  $\phi$  is called *father wavelet*. In this case, a wavelet function  $\psi$ , sometimes called *mother wavelet* can be found directly as follows. Since  $V_1 = V_0 \oplus W_0$ , let  $W_0 = \text{span}\{\sum_{k \in \mathbb{Z}} d_k \psi(x - k), d_k \in \mathbb{R}\}$ . If we can find a function  $\psi \in V_1$  which is orthonormal to  $V_0$  and integer translates of  $\psi$  are orthonormal, then  $\psi$  can be shown to be an orthonormal wavelet. Indeed, we can easily see that  $\psi_{jk} \in V_{j+1}$ . To see  $\psi_{jk}$  are orthonormal, let us consider the inner product of  $\psi_{jk}$  and  $\psi_{mn}$ . Without loss of generality, let us assume that  $j < m$ . Then  $\psi_{jk}$  is in  $V_{j+1} \subset V_m$  and  $\psi_{mn} \in V_m$  which is orthonormal to  $V_{j+1}$  and hence, they are orthogonal to each other. Once  $V_j, j \in \mathbb{Z}$ , are dense in  $L_2(\mathbb{R})$ , it immediately follows that  $\psi$  is an orthonormal wavelet.

In order to find such a function  $\psi \in V_1$ , we take Fourier transform of Eq. (6) to get

$$\widehat{\phi}(\omega) = P(\omega/2) \widehat{\phi}(\omega/2), \quad (8)$$

where  $P(\omega) = \frac{1}{2} \sum_{k \in \mathbb{Z}} p_k e^{-ik\omega}$ . Using Parseval's equality to Eq. (7), we have

$$\begin{aligned} \delta_{0k} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |\widehat{\phi}(\omega)|^2 e^{-ik\omega} d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{n \in \mathbb{Z}} |\widehat{\phi}(\omega + 2n\pi)|^2 e^{-ik\omega} d\omega. \end{aligned}$$

It follows that

**Theorem 5**  $\phi$  is orthonormal if and only if

$$\sum_{n \in \mathbb{Z}} |\widehat{\phi}(\omega + 2n\pi)|^2 = 1.$$

In terms of (8), we have

$$|P(\omega)|^2 + |P(\omega + \pi)|^2 = 1. \quad (9)$$

Since  $\psi \in V_1$ , let us write  $\psi(x) = \sum_{k \in \mathbb{Z}} q_k \phi(2x - k)$ . In terms of Fourier transform, we have  $\widehat{\psi}(\omega) = Q(\omega/2) \widehat{\phi}(\omega/2)$ . The analysis similar to the above shows that if  $\psi(x - k), k \in \mathbb{Z}$  are orthonormal to each other, then

$$|Q(\omega)|^2 + |Q(\omega + \pi)|^2 = 1.$$

Also, a similar analysis of  $\int_{-\infty}^{\infty} \psi(x) \phi(x - k) dx = 0$  for all  $k \in \mathbb{Z}$  implies that

$$P(\omega) \overline{Q(\omega)} + P(\omega + \pi) \overline{Q(\omega + \pi)} = 0.$$

That is, the matrix

$$\begin{bmatrix} P(\omega) & P(\omega + \pi) \\ Q(\omega) & Q(\omega + \pi) \end{bmatrix}$$

is a unitary matrix which is a necessary condition for  $\psi$  to be orthonormal. The solution of  $Q$  is trivial and it is  $Q(\omega) = e^{i\omega} P(\omega + \pi)$ . We remark here that after multiplying them by  $\sqrt{2}$ , the four Laurent polynomials  $P(\omega), P(\omega + \pi), Q(\omega)$ , and  $Q(\omega + \pi)$  form a conjugate quadrature filter bank (see Example 2 in the previous section).

The above discussion explains an excellent approach to construct wavelet functions. We can summarize as follows.

**Theorem 6 (Multiresolution Approximation)** Let  $V_j, j \in \mathbb{Z}$  be a nested sequence of subspaces of  $L_2(\mathbb{R})$ . That is,  $V_j \subset V_{j+1}, \forall j \in \mathbb{Z}$ . Suppose that  $\bigcup_{j \in \mathbb{Z}} V_j$  is dense in  $L_2(\mathbb{R})$  and  $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$ . If there exists a scaling or refinable function  $\phi \in V_0$  whose integer translates span  $V_0$  such that there exist positive constants  $A$  and  $B$

$$A \leq \sum_{k \in \mathbb{Z}} |\hat{\phi}(\omega + 2k\pi)|^2 \leq B,$$

then there exists an orthonormal wavelet function  $\psi \in V_1$  such that  $\{\psi_{jk}, j, k \in \mathbb{Z}\}$  is an orthonormal basis for  $L_2(\mathbb{R})$ .

A proof of this result is provided in [70]. We now explain the usefulness when a nested sequence  $\{V_j, j \in \mathbb{Z}\}$  forms a multiresolution approximation of  $L_2(\mathbb{R})$ . To approximate a function  $f \in L_2(\mathbb{R})$ , we can use functions in  $V_j$  for some integer  $j$ . If functions in  $V_j$  can not approximate  $f$  very well, we look for functions in the next level  $V_{j+1}$  to approximate  $f$  by adding functions in  $W_j$ . We can continue to add functions in  $W_{j+1}, W_{j+2}, \dots$  to get better and better approximation of  $f$ . Thus we can build several levels of approximations of  $f$ . In other word,  $f$  can be approximated in different resolutions.

When  $A \neq B$  in Theorem 6,  $\psi$  is not in general compactly supported. To obtain a compactly supported orthonormal wavelet, we have to study the properties of  $\phi$  to satisfy the necessary and sufficient condition in Theorem 5. There are several equivalent conditions which can be found in [24]. We shall discuss the construction of compactly supported orthonormal wavelets as well as other wavelet functions in later sections.

### Wavelet Decomposition and Reconstruction

Suppose that we have a compactly supported refinable function  $\phi \in L_2(\mathbb{R})$  which generates an MRA of  $L_2(\mathbb{R})$ . Let  $V_j = \text{span}\{\phi_{j,k}, k \in \mathbb{Z}\}$  for all  $j \in \mathbb{Z}$  and  $W_j$  be the orthogonal complement of  $V_j$  in  $V_{j+1}$ , where  $\phi_{jk} = 2^{j/2}\phi(2^j x - k)$  as usual. Also let  $\psi \in W_0$  be a wavelet function which spans  $W_0$  in the sense that  $W_0 = \text{span}\{\psi(x - k), \forall k \in \mathbb{Z}\}$  and

$$A \leq \sum_{k \in \mathbb{Z}} |\hat{\psi}(\omega + 2k\pi)|^2 \leq B$$

for two positive constants  $A$  and  $B$ . Because  $V_0 \subset V_1$ ,  $W_0 \subset V_1$  and  $V_1 = V_0 + W_0$ , we have the following decomposition and reconstruction sequences:

$$\begin{aligned} \phi(x) &= \sum_{k \in \mathbb{Z}} p_k 2^{1/2} \phi(2x - k) \\ \psi(x) &= \sum_{k \in \mathbb{Z}} q_k 2^{1/2} \phi(2x - k) \end{aligned} \quad (10)$$

and

$$2^{1/2} \phi(2x - k) = \sum_{j \in \mathbb{Z}} (a_{2j-k} \phi(x - j) + b_{2j-k} \psi(x - j)), \quad (11)$$

where only finitely many  $p_k, q_k, a_j, b_j$  are nonzero.

Let  $f \in L_2(\mathbb{R})$  be a band-limited signal as explained in Sect. “Definition of Filters”. We first approximate  $f$  by the MRA generated by  $\phi$  at a fine level. That is,

$$A_j f(x) = \sum_{k \in \mathbb{Z}} s_{j,k} \phi_{jk}(x),$$

for some coefficients  $s_{j,k}$ . For example, when  $\phi$  is orthonormal, we may use orthogonal projection,

$$s_{j,k} := \langle f, \phi_{jk} \rangle.$$

As another example, we may choose  $s_{j,k} = f(k/2^j)$  to be the digital samples of  $f$  if  $j$  is sufficiently large such that  $1/2^j < \pi/\omega_0$ , where  $\omega_0$  is the half of the bandwidth of  $f$ . All we need is to make sure that  $A_j f$  converges to  $f$  in  $L_2(\mathbb{R})$  as  $j \rightarrow \infty$ . Now we can decompose  $A_j f$  into

$$A_{j-1} f = \sum_{k \in \mathbb{Z}} s_{j-1,k} \phi_{j-1,k} \in V_{j-1}$$

and

$$D_{j-1} f = \sum_{k \in \mathbb{Z}} d_{j-1,k} \psi_{j-1,k} \in W_{j-1}$$

by using (11) with

$$\begin{aligned} s_{j-1,k} &= \sum_{m \in \mathbb{Z}} s_{j,m} a_{2k-m} \\ d_{j-1,k} &= \sum_{m \in \mathbb{Z}} s_{j,m} b_{2k-m}. \end{aligned}$$

Note that the digital signal  $\{s_{j-1,k}\}$  is computed from  $\{s_{j,k}\}$  by first convolution with a digital filter  $\{a_k\}$  and then downsampling by 2. Similar for  $\{d_{j-1,k}\}$ . We can certainly continue such a decomposition step. That is, we get  $A_j f, A_{j-1} f, \dots, A_{j-\ell} f$  and  $D_{j-1} f, D_{j-2} f, \dots, D_{j-\ell} f$  for some  $\ell \geq 1$ .

On the other hand, we can reconstruct  $A_j f$  back from these  $A_{j-\ell} f, D_{j-\ell}, D_{j-\ell+1}, \dots, D_{j-1} f$ . Indeed, let us say  $\ell = 1$ . Then

$$\begin{aligned} A_{j-1} f &= \sum_{k \in \mathbb{Z}} s_{j-1,k} \phi_{j-1,k} = \sum_{k \in \mathbb{Z}} s_{j-1,k} \sum_{m \in \mathbb{Z}} p_{m-2k} \phi_{j,m} \\ &= \sum_{m \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} s_{j-1,k} p_{m-2k} \phi_{j,m}. \end{aligned}$$

Similar for  $D_{j-1}f$ . Thus,  $A_j f = A_{j-1}f + D_{j-1}f$  implies that

$$s_{j,m} = \sum_{k \in \mathbb{Z}} s_{j-1,k} p_{m-2k} + \sum_{k \in \mathbb{Z}} d_{j-1,k} q_{m-2k}.$$

In terms of signal processing, the digital signal  $\{s_{j,k}\}$  is obtained from  $\{s_{j-1,k}\}$  and  $\{d_{j-1,k}\}$  by first upsampling  $\{s_{j-1,k}\}$  by 2 then convoluting with digital filter  $\{p_k\}$ , secondly upsampling  $\{d_{j-1,k}\}$  by 2 and convoluting with  $\{q_k\}$ , and finally adding the resulting signals together.

The above decomposition/reconstruction procedures are exactly the same as the well-known subband coding of signals described in Sect. “Definition of Filters”. It is clear that the decomposition/reconstruction can be generalized to deal with 2D signals. In other words, 2D images can be decomposed and reconstructed in the same fashion. This decomposition and reconstruction process is very popular in applications such as image/signal denoising, image/signal feature extraction, and image/signal compression. A less popular, but useful idea to use wavelets is to detect singularities from signal by examining the wavelet coefficients  $D_{j-1}f, \dots, D_{j-\ell}f$ . Another less popular, but useful, method is to use wavelets to build more accurate solutions. That is, if we have  $A_j f$  and if we compute  $D_j f$  from  $f$  directly, then we can build up  $A_{j+1}f$  by letting  $A_{j+1}f = A_j f + D_j f$ . In particular, when  $D_j f$  may be computed using the same expense as  $A_j f$ ,  $A_{j+1}f$  can be done in parallel to achieve the efficiency.

### Refinable Functions

As we saw in the previous section, we need a refinable function  $\phi$  to build a multiresolution approximation of  $L_2(\mathbb{R})$ . That is, we want to have

$$\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2)$$

for some Laurant polynomial function  $P(\omega)$ . Fortunately, there are many known functions satisfying the above dilation relation. Fix any integer  $n \geq 1$  and let

$$\widehat{N}_n(\omega) = \left( \frac{1 - e^{i\omega}}{i\omega} \right)^n. \quad (12)$$

This is the well-known *uniform B-spline* function of order  $n$  over integer knot sequence  $\{0, 1, \dots, n\}$ . Since  $1 - e^{i\omega} = (1 + e^{i\omega/2})(1 - e^{i\omega/2})$ , this  $N_n$  is refinable with  $P(\omega) = (1 + e^{i\omega})^n/2^n$ . See [6] and [84] for the properties of B-splines.

To construct other refinable functions, we have to study what kinds of sequence  $\{p_k, k \in \mathbb{Z}\}$  such that

$P(\omega) = \sum_{k \in \mathbb{Z}} p_k e^{ik\omega}$  can generate a refinable function  $\phi$  in the following sense:

$$\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2) = \dots = \prod_{j=1}^{\infty} P(\omega/2^j)\widehat{\phi}(0).$$

It follows that  $P(0)$  must be equal to 1 in order that the above infinite product converges. When  $P(0) = 1$  and  $P(\omega)$  is a Laurent polynomial in the sense that only finitely many  $p_k$  are nonzero,  $\widehat{\phi}(\omega)$  is a continuous function. Furthermore, the distribution  $\phi$  is compactly supported by using Paley–Wiener’s Theorem. It turns out that the condition (9) ensures that  $\widehat{\phi}$  belongs to  $L_2(\mathbb{R})$  (cf. [70]). Hence, when  $P(0) = 1$ ,  $P(\omega)$  is a Laurent polynomial, and  $P(\omega)$  satisfies (9),  $\phi$  is a compactly supported function in  $L_2(\mathbb{R})$ . We still need to check if  $\phi$  satisfies the necessary and sufficient condition for orthonormality in Theorem 5. There is an example that  $P$  satisfies the above mentioned three properties, but  $P$  is not orthonormal (cf. [24]).

If  $\phi$  is orthonormal, we can use the constructive procedure in Sect. “Definition of Filters”, that is, Theorem 6 to find the associated wavelet function. This constitutes a powerful method of construction of univariate orthonormal wavelets. All the construction of various wavelets are based this idea and started with the construction of refinable functions.

Once we have wavelets, we would like to know the smoothness property of the wavelets. Clearly, the smoothness of  $\phi$  determines the smoothness of the associated wavelet function. In terms of Fourier transform,  $\phi \in H^k(\mathbb{R})$  if and only if

$$\int_{-\infty}^{\infty} (1 + |\omega|^2)^k |\widehat{\phi}(\omega)|^2 d\omega < +\infty$$

for some  $k > 0$ . Thus,  $\widehat{\phi}$  needs a decay factor  $O(|\omega|^{-(k+1)})$ . One way to get it is to use the Fourier transform of B-spline of degree  $k$  (or order  $k + 1$ ). Using this factor,

$$\widehat{\phi}(\omega) = \left( \frac{1 - e^{i\omega}}{i\omega} \right)^{k+1} \widetilde{\phi}(\omega)$$

for some function  $\widetilde{\phi} \in L_2(\mathbb{R})$ . In this case,  $\widehat{\phi}(2\omega) = P(\omega)\widehat{\phi}(\omega)$  with mask  $P(\omega) = \left( \frac{1 + e^{i\omega}}{2} \right)^{k+1} \widehat{P}(\omega)$  for some Laurent polynomial  $\widehat{P}(\omega)$ . It turns out that the factor is related to the polynomial reproductivity by the linear combination of integer translates of  $\phi$ . By Strang–Fix conditions ([89] or [60]), a polynomial  $q$  of degree  $\leq k$  can be

expressed by

$$q(x) = \sum_{j \in \mathbb{Z}} c_j(q) \phi(x - j)$$

for some coefficients  $c_j(q)$  if and only if  $\widehat{\phi}$  contains this factor.

### Compactly Supported Orthonormal Wavelets

As we discussed in previous sections, the necessary conditions for  $\phi$  to be an orthonormal scaling function are placed on the mask polynomial  $P(\omega)$ , where we recall that  $\widehat{\phi}(2\omega) = P(\omega)\widehat{\phi}(\omega)$ . They are

- 1)  $P(0) = 1$ ;
- 2)  $|P(\omega)|^2 + |P(\omega + \pi)|^2 = 1$ ;
- 3)  $P(\omega)$  contains a factor  $\left(\frac{1+e^{i\omega}}{2}\right)^k$  for some  $k > 0$ .

Let  $P(\omega) = \left(\frac{1+e^{i\omega}}{2}\right)^k \tilde{p}(\omega)$  and  $|P(\omega)|^2 = |\cos(\omega/2)|^{2k} |\tilde{p}(\omega)|^2$ . For simplicity we write  $x = \cos^2(\omega/2)$  and  $|\tilde{p}(\omega)|^2 = p(1-x)$ . Then the requirement 2) is equal to

$$x^k p(1-x) + (1-x)^k p(x) = 1. \quad (13)$$

It is easy to see that

$$\begin{aligned} 1 &= (1-x+x)^{2k+1} = \sum_{j=0}^k \binom{2k+1}{j} (1-x)^j x^{2k-j} \\ &\quad + \sum_{j=0}^k \binom{2k+1}{2k+1-j} (1-x)^{2k-j} x^j \\ &= x^k \sum_{j=0}^k \binom{2k+1}{j} (1-x)^j x^{k-j} \\ &\quad + (1-x)^k \sum_{j=0}^k \binom{2k+1}{j} (1-x)^{k-j} x^j. \end{aligned} \quad (14)$$

Thus, we may choose  $p(x) = \sum_{j=0}^k \binom{2k+1}{j} (1-x)^{k-j} x^j$ . Note that for  $x = \cos^2(\omega/2) \in [0, 1]$ ,  $p(x) = p(\cos^2(\omega/2)) = p(1/2 + (e^{i\omega} + e^{-i\omega})/4)$  is a Laurent polynomial. Since  $p(x) \geq 0$ , there exists  $\tilde{p}_k(e^{i\omega})$  such that

$$p(x) = |\tilde{p}_k(e^{i\omega})|^2$$

by Fejér–Riesz’s Lemma. It is easy to see that  $p(1-x) = |\tilde{p}_k(-e^{i\omega})|^2$ . Thus, letting

$$P_k(\omega) = \left(\frac{1+e^{i\omega}}{2}\right)^k \tilde{p}_k(e^{i\omega}),$$

we can see that  $P_k$  satisfies 2). Note that  $P_k$  satisfies 1). as well. We define a compactly supported function  $\phi \in L_2(\mathbb{R})$ , in terms of Fourier transform, by

$$\widehat{\phi}_k(\omega) = \prod_{j=1}^{\infty} P_k(\omega/2^j).$$

One can prove (see [24]) that this function  $\phi_k$  is an orthonormal scaling function and generates a multiresolution approximation of  $L_2(\mathbb{R})$  when  $k \geq 1$ . Then we define the associated wavelet  $\psi_k$  by its Fourier transform:

$$\widehat{\psi}_k(\omega) = Q_k(\omega/2) \widehat{\phi}_k(\omega/2),$$

where  $Q_k(\omega) = e^{i\omega} \overline{P_k(\omega + \pi)}$ . As explained in the previous sections,  $\psi_k$  is an orthonormal wavelet. These wavelets are called Daubechies wavelets which were invented in 1988 (cf. [23]).

*Example 3* Let  $k = 1$ . Then  $P_1(\omega) = \frac{1+e^{i\omega}}{2}$  and hence,  $\phi$  is the uniform B-spline of order 1 which is the characteristic function

$$\phi(x) = \begin{cases} 1, & x \in [0, 1) \\ 0, & \text{otherwise} \end{cases}.$$

The associated wavelet is the well-known Haar wavelet:

$$\psi(x) = \begin{cases} 1, & x \in [0, 1/2) \\ -1, & x \in [1/2, 1) \\ 0, & \text{otherwise} \end{cases}.$$

When  $2 \leq k \leq 10$ , the coefficients of  $P(z)$  associated with Daubechies wavelets are given in [24] which can be immediately used for wavelet decomposition and reconstruction.

Next we look at the smoothness of these scaling functions. Unfortunately,  $\phi_k$  does not belong to  $H^k(\mathbb{R})$  since the  $\prod_{j=1}^{\infty} \tilde{p}_k(e^{i\omega/2^j})$  does not belong to  $L_2(\mathbb{R})$ . One has to use a part of the decay factor  $\left(\frac{1+e^{i\omega}}{2}\right)^k$  to ensure  $\prod_{j=1}^{\infty} \left(\frac{1+e^{i\omega/2^j}}{2}\right)^{k-k'} \tilde{p}_k(e^{i\omega/2^j})$  to be in  $L_2(\mathbb{R})$  and thus,  $\phi_k \in H^{k'}(\mathbb{R})$  for some  $k' < k$  (see [24] for  $k'$ ).

### Parameterization of Orthonormal Wavelets

In the previous section we obtained special solutions to the requirements 1), 2), and 3). We now look for general solutions. For convenience, let  $P_n(\omega) = c_0 + c_1 z + \dots + c_{n-1} z^{n-1}$  with  $z^{i\omega}$ . We look for  $P_n(\omega)$  satisfying 1), 2), and 3) with  $k = 1$  discussed in the previous section. That is,  $P_n(0) = 1$ ,  $P_n(\pi) = 0$ , and

$$|P_n(\omega)|^2 + |P_n(\omega + \pi)|^2 = 1, \quad \forall \omega \in \mathbb{R}.$$

In this section, we shall show the solutions  $P_n(\omega)$  for  $n = 4$  and  $n = 6$ . For general  $n \geq 8$  see [59]. As in the previous section, once we have  $P_n$ , we can define a refinable function  $\phi_n \in L_2(\mathbb{R})$  and its associated wavelet function  $\psi_n$  in terms of Fourier transform, by

$$\widehat{\phi}_n(\omega) = \prod_{j=1}^{\infty} P_n(\omega/2^j)$$

and with  $Q_n(\omega) = \overline{zP_n(\omega + \pi)}$  and  $z = e^{i\omega}$ ,

$$\widehat{\psi}_n(\omega) = Q_n(\omega/2)\widehat{\phi}_n(\omega/2).$$

We have to point out that not all the refinable functions  $\phi_n$  are orthonormal. However, the associated wavelet functions generate tight wavelet frames to be defined in a later section.

We first consider  $n = 4$ . Write  $P_4(\omega) = a_0 + b_0z + a_1z^2 + b_1z^3$ .

**Lemma 1**  $P_4(\omega)$  satisfies  $P_4(0) = 1$  and

$$|P_4(\omega)|^2 + |P_4(\omega + \pi)|^2 = 1, \quad \forall \omega \in \mathbb{R}$$

if and only if

$$\begin{aligned} a_0 &= \frac{1}{4} + \frac{1}{2\sqrt{2}} \cos \alpha, & b_0 &= \frac{1}{4} + \frac{1}{2\sqrt{2}} \sin \alpha, \\ a_1 &= \frac{1}{4} - \frac{1}{2\sqrt{2}} \cos \alpha, & b_1 &= \frac{1}{4} - \frac{1}{2\sqrt{2}} \sin \alpha, \end{aligned}$$

for any  $\alpha \in \mathbb{R}$ .

**Example 4** When  $\alpha = \frac{\pi}{4}$ , we get  $P_4(\omega) = \frac{1+z}{2}$ , which is associated with the Haar wavelet.

**Example 5** When  $\alpha = \frac{5\pi}{12}$ , we get  $a_0 = \frac{1+\sqrt{3}}{8}$  and  $a_1 = \frac{3-\sqrt{3}}{8}$  as well as  $b_0 = \frac{3+\sqrt{3}}{8}$ ,  $b_1 = \frac{1-\sqrt{3}}{8}$ . Then  $P_4(\omega)$  is associated with the Daubechies D4 wavelet.

Next, we look for choices of  $\alpha$  where  $P_4(\omega)$  has a second-order vanishing moment, that is,  $P_4(\omega) = \left(\frac{1+z}{2}\right)^2 \tilde{p}(z)$  where  $\tilde{p}(z)$  is some trigonometric polynomial.

In this case  $\frac{d}{dz} P_4(\omega)|_{z=-1} = 0$  which is equivalent to  $\alpha = \frac{5\pi}{12}$  or  $\alpha = \frac{13\pi}{12}$ . Both are associated with Daubechies' D4 wavelet. Thus, D4 is the only member of this family with two vanishing moments.

Next we consider  $P_6(\omega) = a_0 + b_0z + a_1z^2 + b_1z^3 + a_2z^4 + b_2z^5$ .

**Lemma 2**  $P_6(\omega)$  satisfies  $P_6(0) = 1$  and

$$|P_6(\omega)|^2 + |P_6(\omega + \pi)|^2 = 1, \quad \forall \omega \in \mathbb{R}$$

if and only if

$$\begin{aligned} a_0 &= \frac{1}{8} + \frac{1}{4\sqrt{2}} \cos \alpha + \frac{p}{2} \cos \beta \\ a_1 &= \frac{1}{4} - \frac{1}{2\sqrt{2}} \cos \alpha \\ a_2 &= \frac{1}{8} + \frac{1}{4\sqrt{2}} \cos \alpha - \frac{p}{2} \cos \beta \\ b_0 &= \frac{1}{8} + \frac{1}{4\sqrt{2}} \sin \alpha + \frac{p}{2} \sin \beta \\ b_1 &= \frac{1}{4} - \frac{1}{2\sqrt{2}} \sin \alpha \\ b_2 &= \frac{1}{8} + \frac{1}{4\sqrt{2}} \sin \alpha - \frac{p}{2} \sin \beta, \end{aligned}$$

where

$$p = \frac{1}{2} \sqrt{1 + \sin\left(\alpha + \frac{\pi}{4}\right)}$$

for any  $\alpha, \beta \in \mathbb{R}$ .

**Example 6** If  $\alpha = \frac{\pi}{4}$  and  $\beta = \frac{\pi}{4}$ , then  $P_6(\omega)$  is associated with the Haar wavelet.

**Example 7** If  $\alpha = \frac{5\pi}{12}$  and  $\beta = \frac{\pi}{3}$ , then  $P_6(\omega)$  is associated with the Daubechies D4 wavelet.

**Example 8** If

$$\begin{aligned} \cos \alpha &= -\frac{1}{4} \sqrt{8 + \sqrt{15 + 12\sqrt{10}}}, \\ \sin \alpha &= \frac{1}{4} \sqrt{8 - \sqrt{15 + 12\sqrt{10}}}, \\ \cos \beta &= \frac{1}{4} \sqrt{8 - 8\sqrt{-25 + 8\sqrt{10}}}, \\ \sin \beta &= \frac{1}{4} \sqrt{8 + 8\sqrt{-25 + 8\sqrt{10}}}, \end{aligned}$$

then  $P_6(\omega)$  is the filter associated with Daubechies D6 wavelet.

**Example 9** To see when  $P_6(\omega)$  has two vanishing moments, we require  $P'(\pi) = 0$  which implies

$$\sin(\alpha + \pi) = \frac{1}{8 \sin^2(\beta - \pi/4)} - 1.$$

When the ordered pair  $(\alpha, \beta)$  satisfy the above equation,  $P_6(\omega)$  has two vanishing moments.

**Example 10** With  $\alpha = 1.4288992721907328$ ,  $\beta = 1.1071487177940904$ ,  $P_6(\omega)$  is associated with the most smooth length-four filter as given in [24]. When  $\alpha = 1.9886461158096038$  and  $\beta = 1.0934936891036087$ ,  $P_6(\omega)$  is associated with the most smooth length-six filter as given in [24].



## Biorthogonal Wavelets

In this section we explain biorthogonal wavelets. Suppose that  $\phi$  and  $\tilde{\phi}$  are two refinable functions in  $L_2(\mathbb{R})$  which are dual to each other in the following sense:

$$\int_{-\infty}^{\infty} \phi(x-j)\tilde{\phi}(x-k)dx = \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

for all  $j, k \in \mathbb{Z}$ . As usual, define  $\phi_{jk}(x) = 2^{j/2}\phi(2^jx - k)$  and  $\tilde{\phi}_{jk}(x) = 2^{j/2}\tilde{\phi}(2^jx - k)$  for all  $j, k \in \mathbb{Z}$ . Also we let  $V_j = \text{span}\{\phi_{j,k}, k \in \mathbb{Z}\}$  and  $\tilde{V}_j = \text{span}\{\tilde{\phi}_{j,k}, k \in \mathbb{Z}\}$ . We look for  $\psi \in V_1$  and  $\tilde{\psi} \in \tilde{V}_1$  such that

$$\begin{aligned} \langle \phi(\cdot), \tilde{\psi}(\cdot - k) \rangle &= 0, \quad \langle \psi(\cdot), \tilde{\phi}(\cdot - k) \rangle = 0, \\ \langle \psi(\cdot), \tilde{\psi}(\cdot - k) \rangle &= \delta_k, \end{aligned} \quad (16)$$

for all  $k \in \mathbb{Z}$ , where  $\delta_k = 1$  when  $k = 0$  and  $= 0$  for other  $k$ . Then  $\psi$  and  $\tilde{\psi}$  will be a pair of biorthogonal wavelets under the assumptions that both families  $\{\phi_{jk}, j, k \in \mathbb{Z}\}$  and  $\{\tilde{\phi}_{jk}, j, k \in \mathbb{Z}\}$  are multiresolution approximations of  $L_2(\mathbb{R})$ .

Indeed, let  $\psi_{jk}(x) = 2^{j/2}\psi(2^jx - k)$  and  $\tilde{\psi}_{jk}(x) = 2^{j/2}\tilde{\psi}(2^jx - k)$  for all  $j, k \in \mathbb{Z}$ . For any  $j, k$  and  $m, n$  we claim that

$$\int_{-\infty}^{\infty} \psi_{j,k}(x)\tilde{\psi}_{m,n}(x)dx = \delta_{j,m}\delta_{k,n}.$$

For  $j < m$ , we know from (16) that  $\tilde{\psi}_{m,n}$  is orthogonal to  $\phi_{m,k}$  for all  $k$  and hence is orthogonal to  $V_m$  which contains  $\psi_{j+1}$  since  $\psi_{j,\ell} \in V_{j+1} \subset \dots \subset V_m$ . Similar for  $j > m$ . When  $j = m$ , we use (16) again. We now need the following concept:

**Definition 1** A family  $\{\phi_{j,k}, j, k \in \mathbb{Z}\}$  is a Riesz basis for  $L_2(\mathbb{R})$  if

1.  $\phi_{j,k}, j, k \in \mathbb{Z}$  are linearly independent, and
2. there exist two strictly positive constants  $A$  and  $B$  such that for any  $f \in L_2(\mathbb{R})$ ,

$$A\|f\|_2^2 \leq \sum_{j,k} \left| \int_{-\infty}^{\infty} f(x)\phi_{j,k}(x)dt \right|^2 \leq B\|f\|_2^2.$$

Let  $P$  and  $\tilde{P}$  be the mask polynomials associated with  $\phi$  and  $\tilde{\phi}$ . That is,

$$\hat{\phi}(\omega) = P(\omega/2)\hat{\phi}(\omega/2) \quad \text{and} \quad \hat{\tilde{\phi}}(\omega) = \tilde{P}(\omega/2)\hat{\tilde{\phi}}(\omega/2).$$

Then the condition (15) implies

$$P(\omega)\overline{\tilde{P}(\omega)} + P(\omega + \pi)\overline{\tilde{P}(\omega + \pi)} = 1. \quad (17)$$

Let us define the associated wavelet functions  $\psi$  and  $\tilde{\psi}$  in terms of Fourier transform by

$$\hat{\psi}(\omega) = Q(\omega/2)\hat{\phi}(\omega/2) \quad \text{and} \quad \hat{\tilde{\psi}}(\omega) = \tilde{Q}(\omega/2)\hat{\tilde{\phi}}(\omega/2),$$

where  $Q(\omega) = e^{i\omega}\overline{\tilde{P}(\omega + \pi)}$  and  $\tilde{Q}(\omega) = e^{i\omega}\overline{P(\omega + \pi)}$ . The conditions in (16) give

$$\begin{bmatrix} P(\omega) & P(\omega + \pi) \\ Q(\omega) & Q(\omega + \pi) \end{bmatrix} \begin{bmatrix} \tilde{P}(\omega) & \tilde{P}(\omega + \pi) \\ \tilde{Q}(\omega) & \tilde{Q}(\omega + \pi) \end{bmatrix}^* = I_2$$

where  $I_2$  denotes the identity matrix of size  $2 \times 2$ .

In fact, the  $Q$  and  $\tilde{Q}$  so defined as above, one only needs to solve (17). If  $P$  and  $\tilde{P}$  satisfy (17) and if the  $\phi$  and  $\tilde{\phi}$  generate two Riesz bases for  $L_2(\mathbb{R})$ , then one can show that  $\psi$  and  $\tilde{\psi}$  are biorthogonal wavelet functions (cf. [22])

In this setting, for any  $f \in L_2(\mathbb{R})$ , we have

$$f(x) = \sum_{j,k \in \mathbb{Z}} \langle f, \tilde{\psi}_{j,k} \rangle \psi_{j,k}(x) = \sum_{j,k \in \mathbb{Z}} \langle f, \psi_{j,k} \rangle \tilde{\psi}_{j,k}(x).$$

To construct some examples, we begin with the uniform B-spline  $\phi_k$  whose Fourier transform is  $\left(\frac{1-e^{i\omega}}{i\omega}\right)^k$ . It is well-known that  $\phi_k$  is a refinable and generate a multiresolution approximation of  $L_2(\mathbb{R})$ . Let  $P_k(\omega) = \left(\frac{1+e^{i\omega}}{2}\right)^k$  be its mask polynomial. Next we construct a dual  $\phi^{k,n}$  whose mask polynomial

$$\tilde{P}_{k,n}(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^{2n-k} e^{in\omega} \overline{p_n(\sin^2(\omega/2))}$$

where  $p_n(x)$  is a polynomial in  $x$  defined in (14). So that

$$\begin{aligned} & P(\omega)\overline{\tilde{P}_{k,n}(\omega)} + P(\omega + \pi)\overline{\tilde{P}_{k,n}(\omega + \pi)} \\ &= \left(\frac{1+e^{i\omega}}{2}\right)^{2n} e^{-in\omega} p_n(\sin^2(\omega/2)) + \\ & \quad \left(\frac{1-e^{i\omega}}{2}\right)^{2n} e^{-in\omega} (-1)^n p_n(\cos^2(\omega/2)) \\ &= (\cos^2(\omega/2))^n p_n(\sin^2(\omega/2)) \\ & \quad + (\sin^2(\omega/2))^n p_n(\cos^2(\omega/2)) = 1 \end{aligned}$$

by using (14) in a previous section. If we choose  $n$  large enough we can ensure that  $\phi^{k,n} \in L_2(\mathbb{R})$  and satisfies the duality relation (15) with  $\phi_k$  by using the equation above.

There are many other choices of biorthogonal wavelets. In the following we explain a well-known biorthogonal wavelet called CDF 9/7 wavelet (cf. [50]). It was famous for finger print compression employed by FBI in the beginning of 90's. Now it is exclusively used by the state-of-art JPEG2000 for image compression.

**Example 11** Letting  $z = e^{i\omega}$ , we choose

$$P(\omega) = \left(\frac{1+z}{2}\right)^4 (c_0 + c_1 z + (1 - 2(c_0 + c_1))z^2 + c_1 z^3 + c_0 z^4)$$

$$\tilde{P}(\omega) = \left(\frac{1+z}{2}\right)^4 z(\tilde{c}_0 + \tilde{c}_1 z + \tilde{c}_0 z^2).$$

The reason to choose this pattern of coefficients is to make  $\phi$  and  $\tilde{\phi}$  symmetric.

The necessary biorthogonal condition (17) implies that the coefficients  $c_0, c_1$  and  $\tilde{c}_0, \tilde{c}_1$

$$c_0 = \frac{5}{24} + \frac{\sqrt{15}-5}{48}\alpha + \frac{2\sqrt{15}-5}{336}\alpha^2,$$

$$c_1 = -\frac{3}{2} + \frac{6-\sqrt{15}}{12}\alpha - \frac{\sqrt{15}-3}{24}\alpha^2,$$

$$\tilde{c}_0 = -\frac{1}{3} - \frac{1}{12}\alpha + \frac{3\sqrt{15}-11}{168}\alpha^2,$$

$$\tilde{c}_1 = \frac{5}{3} + \frac{\alpha}{6} - \frac{3\sqrt{15}-11}{84}\alpha^2,$$

where  $\alpha = (154 + 42\sqrt{15})^{1/3}$ . This leads to the well-known CDF biorthogonal 9/7 wavelets. Writing  $P(z) = \sum_{i=0}^8 p_i z^i$ ,  $Q(z) = \sum_{i=0}^6 q_i z^i$ ,  $\tilde{P}(z) = \sum_{i=1}^7 \tilde{p}_i z^i$  and  $\tilde{Q}(z) = \sum_{i=-1}^7 \tilde{q}_i z^i$  to be the mask polynomials associated with the CDF 9/7 wavelets, their coefficients are listed in Table 1 which can be directly used for computation.

**Popular Wavelet Families and Filters and Their Use, Table 1**

Numerical values of  $\{p_k\}_{k=0}^8$ ,  $\{q_k\}_{k=0}^6$ ,  $\{\tilde{p}_k\}_{k=1}^7$ , and  $\{\tilde{q}_k\}_{k=-1}^7$

$p_0$	.0534975148216202	$q_0$	.0912717631143501
$p_1$	-.0337282368857499	$q_1$	-.0575435262285002
$p_2$	-.1564465330579805	$q_2$	-.5912717631142501
$p_3$	.5337282368857499	$q_3$	1.1150870524570004
$p_4$	1.2058980364727207	$q_4$	-.5912717631142501
$p_5$	.5337282368857499	$q_5$	-.0575435262285002
$p_6$	-.1564465330579805	$q_6$	.0912717631142501
$p_7$	-.0337282368857499		
$p_8$	.0534975148216202		
		$\tilde{q}_{-1}$	.0534975148216202
$\tilde{p}_0$	.0	$\tilde{q}_0$	.0337282368857499
$\tilde{p}_1$	-.0912717631142501	$\tilde{q}_1$	-.1564465330579805
$\tilde{p}_2$	-.0575435262285002	$\tilde{q}_2$	-.5337282368857499
$\tilde{p}_3$	.5912717631142501	$\tilde{q}_3$	1.2058980364727207
$\tilde{p}_4$	1.1150870524570004	$\tilde{q}_4$	-.5337282368857499
$\tilde{p}_5$	.5912717631142501	$\tilde{q}_5$	-.1564465330579805
$\tilde{p}_6$	-.0575435262285002	$\tilde{q}_6$	.0337282368857499

## Prewavelets

In this section, we first construct a compactly supported pre-wavelet  $\psi_n$  associated with B-spline  $\phi_n = N_n$  of order  $n$ . After that we describe a general approach to construct compactly supported pre-wavelets from any refinable function  $\phi$  who generates a multiresolution approximation of  $L_2(\mathbb{R})$ .

It is known that the B-spline  $N_n$  of order  $n \geq 1$  generates a multiresolution approximation of  $L_2(\mathbb{R})$  (cf. [12]). Let

$$V_0 = \text{span}_{L_2(\mathbb{R})}\{N_n(x-k), \quad k \in \mathbb{Z}\}$$

and  $V_j = \{f(2^j \cdot), \forall f \in V_0\}$ . Since  $N_n$  is a refinable function,  $V_j \subset V_{j+1}$ . Let  $W_j$  be the orthogonal complement of  $V_j$  in  $V_{j+1}$ . If a function  $\psi_n \in W_j$  such that  $W_j = \text{span}_{L_2(\mathbb{R})}\{\psi_n(x-m), m \in \mathbb{Z}\}$ , then  $\psi_n$  is called a prewavelet associated with  $N_n$ .

According to [18], let

$$\psi_n(x) = \frac{1}{2^{n-1}} \sum_{i=0}^{2n-2} (-1)^i N_{2n}(i+1) N_{2n}^{(n)}(2x-j), \quad (18)$$

where  $N_{2n}$  is the uniform B-spline of order  $2n$ . Here  $N_{2n}^{(n)}$  denotes the  $n$ th derivative of  $N_{2n}(x)$  which is a linear combination of B-splines  $N_n(x-k)$  for finitely many  $k$ . In fact,

$$N_{2n}^{(n)}(x) = \sum_{i=0}^n (-1)^i \binom{n}{i} N_n(x-j).$$

It follows that

$$\psi_n(x) = \sum_{i=0}^{3n-2} \frac{(-1)^i}{2^{n-1}} \cdot \sum_{k=0}^n \binom{n}{k} N_{2n}(i-k+1) N_n(2x-i) \in V_1. \quad (19)$$

We can verify that  $\psi_n$  is a pre-wavelet associated with  $N_n$  (cf. [18]).

**Theorem 7** The spline function  $\psi_n(x)$  defined in (18) with support  $[0, 2n-1]$  is a pre-wavelet associated with the B-spline  $N_n(t)$  of order  $n$ .

Let us determine two sequences  $\{g_n\}$  and  $\{h_n\}$  such that

$$N_n(2x-k) = \sum_{m \in \mathbb{Z}} g_{k-2m} N_n(x-m) + \sum_{m \in \mathbb{Z}} h_{k-2m} \psi_n(x-m) \quad (20)$$

for all  $k \in \mathbb{Z}$ .

Let  $G$  and  $H$  be the  $z$ -transform of the sequences  $\{g_i\}_{i \in \mathbb{Z}}$  and  $\{h_i\}_{i \in \mathbb{Z}}$ . That is,

$$G(z) = \sum_{m \in \mathbb{Z}} g_m z^{-m} \quad \text{and} \quad H(z) = \sum_{m \in \mathbb{Z}} h_m z^{-m}.$$

As we know from a previous section,

$$\widehat{N}_n(\omega) = \frac{1}{2} P(\omega) \widehat{N}_n(\omega/2).$$

We find

$$\begin{aligned} \widehat{\psi}_n(\omega) &= \int_{-\infty}^{\infty} \psi_n(x) e^{-ix\omega} dx \\ &= 2^{-n+1} \sum_{i=0}^{2n-2} (-1)^i N_{2n}(i+1) \\ &\quad \int_{-\infty}^{\infty} N_{2n}^{(n)}(2x-i) e^{-ix\omega} dx \\ &= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) e^{-ij\omega/2} \\ &\quad \int_{-\infty}^{\infty} N_{2n}^{(n)}(x) e^{-ix\omega/2} dx \\ &= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) e^{-ij\omega/2} (i\omega/2)^n \\ &\quad \int_{-\infty}^{\infty} N_{2n}(x) e^{-ix\omega/2} dx \\ &= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) e^{-ij\omega/2} \\ &\quad (1 - e^{-i\omega/2})^n \hat{N}_n(\omega/2) \\ &= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) z^j (1-z)^n \hat{N}_n(\omega/2) \\ &= \frac{1}{2} Q(z) \widehat{N}_n(\omega/2) \end{aligned}$$

where  $z = e^{-i\omega/2}$  and

$$Q(z) = 2^{-n+1} (1-z)^n \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) z^j.$$

Note that the Fourier transform of (20) and the orthogonality between  $W_0$  and  $V_0$  can be recast in the following

matrix format:

$$\begin{cases} P(z)G(z) + Q(z)H(z) &= 2 \\ P(-z)G(z) + Q(-z)H(z) &= 0. \end{cases}$$

The solution are

$$\begin{aligned} G(z) &= \frac{2Q(-z)}{P(z)Q(-z) - P(-z)Q(z)}; \\ H(z) &= \frac{-2P(-z)}{P(z)Q(-z) - P(-z)Q(z)}. \end{aligned}$$

We have to verify that  $P(z)Q(-z) - P(-z)Q(z) \neq 0$  for all  $z$ . This follows easily from the following (cf. [18])

**Lemma 3**

$$P(z)Q(-z) - P(-z)Q(z) = \frac{2^{2n} z \tau_{2n-1}(z^2)}{2^{2n-2} (2n-1)!}.$$

where  $\tau_{2n-1}$  is the Euler–Frobenius polynomial which is defined by

$$\tau_n(t) := n! \sum_{i=0}^{n-1} N_{n+1}(i+1) t^i, \quad \forall n \geq 0,$$

which is never zero for  $z$  (cf. [83]).

Therefore,  $\phi$  is a pre-wavelet associated with  $N_n(x)$ . It follows from its definition (19) that  $\psi_n$  is of compact support.

**Example 12** Consider  $m = 1$ . Then

$$N_1(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi_1(x) = N_2(1)N'_2(2x) = N'_2(2x) = \begin{cases} 1 & x \in [0, \frac{1}{2}] \\ -1 & x \in [\frac{1}{2}, 1] \end{cases}$$

which is the Haar wavelet.

**Example 13** Consider  $m = 2$ .

$$N_2(x) = \begin{cases} x, & x \in [0, 1] \\ 2-x, & x \in [1, 2] \\ 0, & \text{otherwise} \end{cases}.$$

Then

$$\begin{aligned} \psi_2(x) &= \frac{1}{2} \sum_{j=0}^2 N_4(j+1) N_4^{(2)}(2x-j) \\ &= \frac{1}{12} (N_4^{(2)}(2x) - 4N_4^{(2)}(2x-1) + N_4^{(2)}(2x-2)) \\ &= \frac{1}{12} (N_2(2x) - 6N_2(2x-1) + 10N_2(2x-2) \\ &\quad - 6N_2(2x-3) + N_2(2x-4)). \end{aligned}$$

We remark that both  $G(z)$  and  $H(z)$  are rational filters. As pointed in Sect. “Definition of Filters”, these are realizable (recursive) filters which can be used for digital signal reconstruction procedure. On the other hand, the zeros  $\lambda_1, \dots, \lambda_{n-1}$  of  $\tau_n(z)$  are located in  $(-\infty, 0)$  and satisfy  $\lambda_{n-1}\lambda_1 = \lambda_{n-2}\lambda_2 = \dots = 1$  (cf. [83]). Thus,  $G$  and  $H$  are not stable IIR filters.

We now consider a new constructive method for finding compactly supported prewavelet function for any given refinable function  $\phi$  which generates a multiresolution approximation of  $L_2(\mathbb{R})$ . Let  $V_j = \text{span}_{L_2(\mathbb{R})}\{\phi(2^j x - k), k \in \mathbb{Z}\}$  for all  $j \in \mathbb{Z}$ . We are looking for compactly supported functions  $\psi$  in  $V_1$  such that

$$V_1 = V_0 \oplus W_0,$$

where  $W_0$  is the closure of the linear span of integer translates of  $\psi(2^j x - m)$ ,  $m \in \mathbb{Z}$  and  $\psi(\cdot - m)$ ,  $m \in \mathbb{Z}$  form a stable basis for  $W_0$ .

To do so, we first introduce a function

$$\Phi(z) := \sum_{m \in \mathbb{Z}} \langle \phi(x), \phi(x - m) \rangle z^m,$$

where  $\langle f, g \rangle$  stands for the standard inner product for  $L_2(\mathbb{R})$ . This function  $\Phi$  may be called the generalized Euler–Frobenius polynomial.

Next we need a necessary and sufficient condition for the orthogonality. Writing

$$g(x) = \sum_{m \in \mathbb{Z}} c_m 2^{1/2} \phi(2x - m) \in V_1,$$

to be a general function in  $V_1$  and

$$G(z) = \frac{1}{\sqrt{2}} \sum_{m \in \mathbb{Z}} c_m z^m,$$

i. e., the Fourier transform  $\widehat{g}(\omega) = G(e^{i\omega/2})\widehat{\phi}(\omega/2)$ , we let

$$\mathcal{G} = \text{closure}_{L_2(\mathbb{R})}\{g(x - m), m \in \mathbb{Z}\}$$

be the closure of the linear span of integer translates of  $g$ . Then we have the following (see, e. g., [49] for a proof)

**Theorem 8** *Let  $P(z)$  be the mask polynomial of  $\phi$ . That is,  $\widehat{\phi}(\omega) = P(e^{i\omega/2})\widehat{\phi}(\omega/2)$ . Then  $\mathcal{G}$  is orthogonal to  $V_0$  if and only if*

$$G(z)\overline{P(z)}\Phi(z) + G(-z)\overline{P(-z)}\Phi(-z) = 0.$$

Our first step is to construct compactly supported  $g_k \in V_1$ ,  $k = 1, 2$  such that the closure  $\mathcal{G}_k$  of the linear span of integer translates  $g_k$  is orthogonal to  $V_0$  for

$k = 1$  and  $k = 2$  and  $V_1 = V_0 \oplus (\mathcal{G}_1 + \mathcal{G}_2)$ . The second step is to find a sufficient condition that one of them can be written in terms of integer translates of the other. For example,  $\mathcal{G}_2$  is contained in  $\mathcal{G}_1$ . In this case we will have  $V_1 = V_j \oplus \mathcal{G}_1$ . To be more precise, we suppose that  $g_k \in V_1$  satisfy

$$g_k(x - m) \perp V_0, \quad m \in \mathbb{Z}$$

for  $k = 1, 2$  and

$$\begin{aligned} 2^{1/2}\phi(2x) &= \sum_{m \in \mathbb{Z}} (a_{1,m}\phi(x - m) + b_{1,m}g_1(x - m)) \\ 2^{1/2}\phi(2x - 1) &= \sum_{m \in \mathbb{Z}} (a_{2,m}\phi_j(x - m) + b_{2,m}g_2(x - m)). \end{aligned}$$

In terms of Fourier transform, the above equations can be rewritten as

$$\begin{aligned} \frac{1}{2^{1/2}}\widehat{\phi}\left(\frac{\omega}{2}\right) &= A_1(\omega)\widehat{\phi}(\omega) + B_1(\omega)\widehat{g}_1(\omega) \\ &= A_1(\omega)P\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) \\ &\quad + B_1(\omega)G_1\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) \end{aligned}$$

and

$$\begin{aligned} \frac{1}{2^{1/2}}e^{i\omega/2}\widehat{\phi}\left(\frac{\omega}{2}\right) &= A_2(\omega)\widehat{\phi}(\omega) + B_2(\omega)\widehat{g}_1(\omega) \\ &= A_2(\omega)P\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) \\ &\quad + B_2(\omega)G_2\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right), \end{aligned}$$

where  $A_k(\omega) = \sum_{m \in \mathbb{Z}} a_{k,m}e^{im\omega}$  and  $B_k(\omega) = \sum_{m \in \mathbb{Z}} b_{k,m}e^{im\omega}$ . Here, we have abused the notation of  $G_k$ , that is, we use  $G_k(\omega)$  instead of  $G_k(z)$  with  $z = e^{i\omega}$  just for convenience.

It follows that

$$\begin{aligned} A_1(2\omega)P(\omega) + B_1(2\omega)G_1(\omega) &= 1, \\ A_2(2\omega)P(\omega) + B_2(2\omega)G_2(\omega) &= e^{i\omega}. \end{aligned}$$

Using Theorem 8, the solution of  $A_k$ ,  $B_k$  and  $G_k$  can be easily found. Indeed, let  $E$  be the operator which maps a Laurent polynomial  $f$  into a Laurent polynomial  $E(f)$  which contains all the even index terms of  $f$ . That is,  $E(f) = (f(z) + f(-z))/2$ . One simple property of  $E$  is  $E(f(z)) = E(f(-z))$ .

**Theorem 9** *Suppose that  $E(P(z)\overline{P(z)}(z)) \neq 0$  for all  $z$  with  $|z| = 1$ . Let*

$$\begin{aligned} A_1(2\omega) &:= \frac{E(\overline{P(z)}\Phi(z))}{E(P(z)\overline{P(z)}\Phi(z))}, \\ B_1(2\omega) &:= \frac{1}{E(P(\omega)\overline{P(\omega)}\Phi(z))}, \\ G_1(\omega) &:= E(P(\omega)\overline{P(\omega)}\Phi(z)) - E(\overline{P(\omega)}\Phi(z))P(\omega) \end{aligned}$$

and

$$\begin{aligned} A_2(2\omega) &:= \frac{E(e^{i\omega} \overline{P(\omega)} \Phi(z))}{E(P(\omega) \overline{P(\omega)} \Phi(z))}, \\ B_2(2\omega) &:= \frac{1}{E(P(\omega) \overline{P(\omega)} \Phi(z))}, \\ G_2(\omega) &:= E(P(\omega) \overline{P(\omega)} \Phi(z)) e^{i\omega} \\ &\quad - E(e^{i\omega} \overline{P(\omega)} \Phi(z)) P(\omega). \end{aligned}$$

Then  $\mathcal{G}_k$  is orthogonal to  $V_0$  for all  $k = 1, 2$  and

$$V_1 = V_0 \oplus (\mathcal{G}_1 + \mathcal{G}_2).$$

*Proof* Using the assumption of Theorem 9, we know that  $A_k, B_k$  are well-defined. It is clear that  $V_1$  is the direct sum of  $V_0$  and  $\mathcal{G}_k, k = 1, 2$ . To see  $\mathcal{G}_k$  is orthogonal to  $V_0$ , we use Theorem 8 to see  $E(G_k(z) \overline{P(z)} \Phi(z)) = 0$ . Since  $B_k(2\omega) \neq 0$  and

$$\begin{aligned} E(B_k(2\omega) G_k(z) \overline{P(z)} \Phi(z)) \\ = B_k(2\omega) E(G_k(z) \overline{P(z)} \Phi(z)), \end{aligned}$$

we may consider

$$\begin{aligned} E(B_1(2\omega) G_1(z) \overline{P(z)} \Phi(z)) \\ = E((1 - A_1(2\omega) P(z)) \overline{P(z)} \Phi(z)) \\ = E(\overline{P(z)} \Phi(z)) - A_1(2\omega) E(P(z) \overline{P(z)} \Phi(z)) = 0 \end{aligned}$$

by the construction of  $A_1$ . Similar for the second equation. This completes the proof.  $\square$

Let us make a remark on  $E(P(\omega) \overline{P(\omega)} \Phi(z))$ . The following result is known (cf. [12]).

**Lemma 4**

$$E(P(\omega) \overline{P(\omega)} \Phi(z)) = \frac{1}{2} \Phi(z^2).$$

Next we show that  $g_k, k = 1, 2$  are linearly dependent if  $P$  satisfies another condition. Let us write  $P$  in its polyphase form, i. e.,

$$P(z) = P_0(z^2) + z P_1(z^2).$$

**Theorem 10** Suppose that  $P_0(z)$  is not zero for any  $z$  with  $|z| = 1$ . Then there exist non-zero coefficients  $f_m$ 's such that

$$g_1(x) = \sum_{m \in \mathbb{Z}} f_m g_2(x - m).$$

Furthermore, the integer translates of  $g_2$  form a Riesz basis for  $W_0 := V_1 \ominus V_0$  which is the closure of the linear span of  $g_2(2^j \cdot -m), m \in \mathbb{Z}$ . If  $P_1(z) \neq 0$  for all  $z$ , then

$$g_2(x) = \sum_{m \in \mathbb{Z}} f'_{j,m} g_1(x - m)$$

for some coefficients  $f'_m$  and the integer translates of  $g_1$  form a Riesz basis for  $W_0$ .

Without loss of generality, we may assume that  $P_1$  is not zero. The letting  $\psi = g_1$  and  $W_0 := \mathcal{G}_1, \psi$  is a prewavelet associated with  $\phi$ . It is easy to see that  $\psi$  is compactly supported if  $\phi$  is since  $\psi = g_1$  is a finitely linear combination of integer translates of  $\phi$  because

$$\widehat{g}_1(\omega) = G_1(\omega) \widehat{\phi}(\omega)$$

with Laurent polynomial  $G_1(\omega)$ . Let  $\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$ . We can verify that all these  $\psi_{jk}$  form a Riesz basis for  $L_2(\mathbb{R})$  (cf. [49]). We have

**Theorem 11** Suppose that  $\phi$  is a compactly supported refinable function generating an MRA for  $L_2(\mathbb{R})$ . Denote  $\widehat{\phi}(\omega) = P(\omega/2) \widehat{\phi}(\omega/2)$ . Let  $\Phi$  be the generalized Euler-Frobenius polynomial associated with  $\phi$ . Suppose that

$$E(P(\omega) \overline{P(\omega)} \Phi(\omega)) \neq 0, \quad \forall e^{i\omega}$$

and suppose that at least one of the two polyphases of  $P$  is not zero, i. e.,  $P_0(\omega) \neq 0$  or  $P_1(\omega) \neq 0$  for all  $\omega \in [0, 2\pi]$ . Then there exist a compactly supported function  $\psi$  such that the closure  $W_0$  of the linear span of integer translates  $\psi(x - m), m \in \mathbb{Z}$  is orthogonal to  $V_0, V_1 = V_0 \oplus W_0$  and the integer translates of  $\psi$  form a Riesz basis for  $W_0$ . All of them forms a Riesz basis for  $L_2(\mathbb{R})$ .

Let us use B-splines to give some examples of prewavelets for  $L_2(\mathbb{R})$ . Recall B-splines from (12). Fix an integer  $n > 0$ . Let  $\phi = N_n$  the  $n$ -th order B-spline. It is easy to see

$$\widehat{N}_n(\omega) = P^{(n)}(\omega/2) \widehat{N}_n(\omega/2)$$

$$\text{with } P^{(n)}(\omega) = \left( \frac{1 + e^{-i\omega}}{2} \right)^n.$$

It is clear that

$$\begin{aligned} P^{(n)}(\omega) - P^{(n)}(\omega + \pi) \\ = \left( \frac{1 + e^{-i\omega}}{2} \right)^n - \left( \frac{1 - e^{-i\omega}}{2} \right)^n \neq 0 \end{aligned}$$

for any  $\omega \in [0, 2\pi]$ . It follows that the polyphase  $P_1^{(n)}$  associated with  $\phi$  is never zero.

Next by Lemma 4, we know

$$E(P^{(n)}(\omega) \overline{P^{(n)}(\omega)} \Phi(\omega)) = \frac{1}{2} \Phi(2\omega).$$



By Poisson summation formula,

$$\begin{aligned}\Phi(\omega) &= \sum_{m \in \mathbb{Z}} \langle N_n(\cdot), N_n(\cdot - m) \rangle e^{-im\omega} \\ &= \sum_{m \in \mathbb{Z}} |\widehat{N}_n(\omega + 2m\pi)|^2\end{aligned}$$

which is never zero (cf. [83]). That is, the conditions in Theorem 11 are satisfied.

The above discussions verify that all B-spline functions can be used to construct prewavelets for  $L_2(\mathbb{R})$ . Note that our B-spline prewavelets have a larger support than those constructed in the first half of this section. The purpose of the examples is to show the detail of the constructive procedure. The advantages of the second construction are (1). it enables us to construct prewavelets from any refinable functions and (2). the construction can be easily generalized to the multivariate setting to be discussed later. Note that the first construction has no multivariate generalization.

**Example 14** Consider linear B-spline  $N_2$  with  $P^{(2)}(z) = (1 + z)^2/4$ . It is easy to see that

$$\Phi(z) = \frac{1}{6}z^{-1} + \frac{4}{6} + \frac{1}{6}z.$$

Indeed, by using the symmetric property of B-splines, i. e.,  $N_2(x) = N_2(2 - x)$ . It is easy to see

$$\Phi(z) = \sum_{m \in \mathbb{Z}} N_4(2 + m)z^m = \frac{1}{24z^2}(10z^2 + z^4 + 1).$$

Thus, we know  $E(P^{(2)}(z)P^{(2)}(1/z)\Phi(z)) = \frac{1}{2}\Phi(z^2) \neq 0$ . Using the formulas in Theorem 9, we have

$$G(z) = \frac{1}{96z^2}(z^6 - 6z^5 + 11z^4 - 12z^3 + 11z^2 - 6z + 1).$$

That is, the prewavelet associated with linear B-spline  $\phi := N_2$  is

$$\begin{aligned}\psi(x) &= \frac{1}{96}(\phi(2x + 2) - 6\phi(2x + 1) + 11\phi(2x) \\ &\quad - 12\phi(2x - 1) + 11\phi(2x - 2) - 6\phi(2x - 3) \\ &\quad + \phi(2x - 4)).\end{aligned}$$

**Example 15** Consider cubic B-spline  $N_4$  with  $P^{(4)}(z) = (1 + z)^4/16$ . We have

$$\begin{aligned}\Phi(z) &= \frac{1}{5040z^3} + \frac{1}{42z^2} + \frac{397}{1680z} + \frac{151}{315} \\ &\quad + \frac{397}{1680}z + \frac{1}{42}z^2 + \frac{1}{5040}z^3.\end{aligned}$$

Thus,  $E(P^{(2)}(z)P^{(2)}(1/z)\Phi(z)) = \frac{1}{2}\Phi(z^2) \neq 0$ .

$$\begin{aligned}E(zP^{(4)}(1/z)\Phi(z)) &= \frac{1}{80640z^6}(18482z^4 + z^{10} + 18482z^6 \\ &\quad + 1677z^2 + 1 + 1677z^8).\end{aligned}$$

Thus, by using the formula in Theorem 9,  $G(z) = \sum_{k=-6}^8 g_k z^k$  with coefficients  $g_k$  as follows:

$$\begin{aligned}g_{-6} &= \frac{-1}{1290240}, & g_{-5} &= \frac{31}{322560}, & g_{-4} &= \frac{-187}{143360} \\ g_{-3} &= \frac{1081}{161280}, & g_{-2} &= \frac{-1903}{86016}, & g_{-1} &= \frac{17953}{322560}, \\ g_0 &= \frac{-131051}{1290240}, & g_1 &= \frac{1441}{11520}, & g_2 &= \frac{-131051}{1290240} \\ g_3 &= \frac{17953}{322560}, & g_4 &= \frac{-1903}{86016}, & g_5 &= \frac{1081}{161280}, \\ g_6 &= \frac{-187}{143360}, & g_7 &= \frac{3}{322560}, & g_8 &= \frac{-1}{1290240}.\end{aligned}$$

That is, the prewavelet associated with cubic B-spline  $\phi := N_4$  is

$$\begin{aligned}\psi(x) &= \frac{1}{1290240}(\phi(2x + 6) - 124\phi(2x + 5) \\ &\quad + 1683\phi(2x + 4) - 8648\phi(2x + 3) \\ &\quad + 28545\phi(2x + 2) - 71812\phi(2x + 1) \\ &\quad + 131051\phi(2x) - 161392\phi(2x - 1) \\ &\quad + 131051\phi(2x - 2) - 71812\phi(2x - 3) \\ &\quad + 28545\phi(2x - 4) - 8648\phi(2x - 5) \\ &\quad + 1683\phi(2x - 6) - 124\phi(2x - 7) + \phi(2x - 8)).\end{aligned}$$

We can easily verify that  $\psi$  is orthogonal to the integer translates of  $\phi$  using the computer program MAPLE.

## Tight Wavelet Frames

In this section we consider the tight wavelet frames. First of all, let us introduce some notation and explain their usefulness for representation of functions in  $L_2(\mathbb{R})$ . We will show the striking feature of tight wavelet frames: their representation of any function  $f \in L_2(\mathbb{R})$  is just like that by using an orthonormal wavelet basis although they can contain redundancy. Also, the representation is most economic in the sense that the  $\ell^2$  norm of the coefficients in the representation is the same as the  $L_2$  norm of the function.

**Definition 2** A family of functions  $\{\phi_j, j \in J\}$  in  $L_2(\mathbb{R})$  is called a *frame*, if there exist constants  $A > 0, B > 0$  such that

$$A\|f\|^2 \leq \sum_{j \in J} |\langle f, \phi_j \rangle|^2 \leq B\|f\|^2, \quad \forall f \in L_2(\mathbb{R}),$$

where  $\|f\|$  denotes the norm of  $f$  in  $L_2(\mathbb{R})$ . If  $A = B$ ,  $\{\phi_j\}_{j \in J}$  is called *tight wavelet frame*.

Thus for a tight wavelet frame we may normalize so that  $A = B = 1$ . In this case, we will have

$$\|f\|^2 = \sum_{j \in J} |\langle f, \phi_j \rangle|^2, \quad \forall f \in L_2(\mathbb{R}). \quad (21)$$

Thus for any  $f$  and  $g$  in  $L_2(\mathbb{R})$ , we have

$$\begin{aligned} \|f + g\|^2 &= \sum_{j \in J} |\langle f + g, \phi_j \rangle|^2, \\ \|f - g\|^2 &= \sum_{j \in J} |\langle f - g, \phi_j \rangle|^2. \end{aligned}$$

Differencing the two equations, we have

$$4\langle f, g \rangle = 4 \sum_{j \in J} \langle f, \phi_j \rangle \langle \phi_j, g \rangle \quad (22)$$

for all  $g \in L_2(\mathbb{R})$ . That is,

$$f = \sum_{j \in J} \langle f, \phi_j \rangle \phi_j, \quad \text{weakly} \quad \forall f \in L_2(\mathbb{R}). \quad (23)$$

Thus, a tight wavelet frame can represent any  $f \in L_2(\mathbb{R})$  just like an orthonormal basis. Note that the representation (23) is simply derived from (21) and the technique is called polarization.

Compared to the orthonormal wavelet basis representation of functions in  $L_2(\mathbb{R})$ , a tight wavelet frame needs no orthonormality nor linear independence among the functions  $\phi_j$ ,  $j \in J$ . In fact, it allows redundancy in  $\phi_j$ ,  $j \in J$ . Thus, we have more degrees of freedom to construct tight wavelet frames. In the following we shall present three methods.

In general we shall construct a finitely many compactly supported functions  $\psi_\ell \in L_2(\mathbb{R})$  such that

$$\begin{aligned} \Lambda(\psi_\ell, \ell = 1, \dots, r) &= \{\psi_{\ell,j,k}(x) := 2^{j/2} \psi_\ell(2x - k), \\ &\quad j, k \in \mathbb{Z}, \quad \ell = 1, 2, \dots, r\} \end{aligned}$$

is a tight wavelet frame.

We start with a refinable function  $\phi \in L_2(\mathbb{R})$ . Let  $P(\omega)$  be the mask polynomial of  $\phi$ . That is,

$$\widehat{\phi}(\omega) = P(\omega/2) \widehat{\phi}(\omega/2).$$

Suppose that  $\phi$  generates a multiresolution approximation of  $L_2(\mathbb{R})$ . For Laurent polynomials  $Q_\ell(\omega)$ ,  $\ell = 1, \dots, r$ , let  $\psi_\ell$  be a function in  $L_2(\mathbb{R})$  defined in terms of Fourier transform by

$$\widehat{\psi}_\ell(\omega) = Q_\ell(\omega/2) \widehat{\phi}(\omega/2), \quad \ell = 1, \dots, r.$$

The following condition is called Unitary Extension Principle (UEP) developed in [76].

$$P(\omega) \overline{P(\omega + v)} + \sum_{\ell=1}^r Q_\ell(\omega) \overline{Q_\ell(\omega + v)} = \begin{cases} 1 & \text{if } v = 0, \\ 0 & \text{if } v = \pi. \end{cases} \quad (24)$$

**Theorem 12** Suppose that  $\phi \in L_2(\mathbb{R})$  is a compactly supported continuous function with Hölder continuity  $\alpha > 0$  and suppose that  $\phi$  generates a multiresolution approximation of  $L_2(\mathbb{R})$ . Assume that there exist Laurent polynomials  $Q_\ell$ ,  $\ell = 1, \dots, r$  such that  $P$  and  $Q_\ell$  satisfy the UEP condition. Then defining  $\psi_\ell$  as above,  $\Lambda(\psi_\ell, \ell = 1, \dots, r)$  is a tight wavelet frame for  $L_2(\mathbb{R})$ .

The proof of the above theorem can be derived based on the following three lemmas.

**Lemma 5** Let  $\phi \in L_2(\mathbb{R})$ . Suppose that  $\widehat{\phi}(0) = 1$  and that for some constant  $B > 0$ ,

$$\sum_{m \in 2\pi\mathbb{Z}} |\widehat{\phi}(\omega + m)|^2 \leq B < +\infty, \quad \text{a. e., } \omega \in \mathbb{R}.$$

Define

$$\beta_j(f, \omega) = 2^{j/2} \sum_{m \in 2\pi\mathbb{Z}} \widehat{f}(2^j(\omega + m)) \overline{\widehat{\phi}(\omega + m)}$$

for any fixed  $f \in L_2(\mathbb{R})$  and  $j \in \mathbb{Z}$ . Then

$$\lim_{j \rightarrow +\infty} \int_{[0, 2\pi]} |\beta_j(f, \omega)|^2 d\omega = \|f\|^2 \quad \text{and}$$

$$\lim_{j \rightarrow -\infty} \int_{[0, 2\pi]} |\beta_j(f, \omega)|^2 d\omega = 0.$$

**Lemma 6** For the refinable function  $\phi$  satisfying the same condition in Lemma 5 and  $\psi_\ell$  defined above, we have the following equations:

$$\begin{aligned} \sum_{k \in \mathbb{Z}} |\langle f, \phi_{j,k} \rangle|^2 &= \int_{[0, 2\pi]} |\beta_j(f, \omega)|^2 d\omega \\ \sum_{k \in \mathbb{Z}} |\langle f, \phi_{j-1,k} \rangle|^2 &= \frac{1}{2} \int_{[0, 2\pi]} \left| \sum_{v \in \{0, 1\}\pi} \overline{P\left(\frac{\omega}{2} + v\right)} \right. \\ &\quad \left. \beta_j\left(f, \frac{\omega}{2} + v\right) \right|^2 d\omega \\ \sum_{k \in \mathbb{Z}} |\langle f, \psi_{j,k}^\ell \rangle|^2 &= \frac{1}{2} \int_{[0, 2\pi]} \left| \sum_{v \in \{0, 1\}\pi} \overline{Q_\ell\left(\frac{\omega}{2} + v\right)} \right. \\ &\quad \left. \beta_j\left(f, \frac{\omega}{2} + v\right) \right|^2 d\omega \end{aligned}$$

where  $\phi_{j,k}(\cdot) = 2^{j/2}\phi(2^j \cdot - k)$ ,  $\psi_{\ell,j,k}(\cdot) = 2^{j/2}\psi_{\ell}(2^j \cdot - k)$ .

**Lemma 7** Suppose that we can find  $Q_{\ell}$ ,  $\ell = 1, \dots, r$  satisfying (24). Let  $\psi_{\ell}$  be the functions defined by its Fourier transform using  $\widehat{\psi_{\ell}}(\omega) = Q_{\ell}(\omega/2)\widehat{\phi}(\omega/2)$ . Then the collection of functions  $\{\psi_{\ell,j,k} : \ell = 1, \dots, r, j, k \in \mathbb{Z}\}$  is a tight wavelet frame.

We now rewrite the UEP in a matrix format.

**Lemma 8** Let  $P = [P(\omega), P(\omega + \pi)]$  be a vector and  $Q = [Q_{\ell}(\omega), Q_{\ell}(\omega + \pi)]_{\ell=1,\dots,r}$  be a matrix of size  $r \times 2$ . Then the UEP condition in (24) is equivalent to

$$Q^*Q = I_2 - PP^*. \quad (25)$$

Here  $I_2$  is the identity matrix of size  $2 \times 2$ .

Multiplying  $P^*$  from the left of (25) and  $P$  from the right, we immediately get

$$P^*P - (P^*P)^2 \geq 0.$$

It follows that

$$P^*P \leq 1 \quad \text{or} \quad |P(\omega)|^2 + |P(\omega + \pi)|^2 \leq 1. \quad (26)$$

That is,  $P$  is necessary to satisfy the above (26) in order to have the associated tight wavelet frame. On the other hand, if a mask polynomial  $P$  satisfies (26), we now show that there is a set of frame generators  $\psi_{\ell}$ ,  $\ell = 1, 2, 3$  such that their dilations and integer translates form a tight wavelet frame for  $L_2(\mathbb{R})$ . That is, the condition (26) is necessary and sufficient to have a tight wavelet frame associated with the refinable function  $\phi$  (cf. [14] and [72]).

Indeed, let  $P(\omega) = P_1(2\omega) + zP_2(2\omega)$  be the poly-phase format of  $P$ , where  $z = e^{i\omega}$ . Since

$$|P(\omega)|^2 + |P(\omega + \pi)|^2 = 2|P_1(2\omega)|^2 + 2|P_2(2\omega)|^2 \leq 1, \quad (27)$$

we let  $P_3(\omega)$  be a polynomial in  $z$  such that

$$2|P_1(2\omega)|^2 + 2|P_2(2\omega)|^2 + |P_3(2\omega)|^2 = 1. \quad (28)$$

This can be done using Riesz–Fejér factorization. We now define a matrix of  $3 \times 3$  by

$$\widetilde{Q}(\omega) = I_3 - \widetilde{P}(\omega)\widetilde{P}(\omega)^*$$

with  $\widetilde{P}(\omega) = [\sqrt{2}P_1(\omega), \sqrt{2}P_2(\omega), P_3(\omega)]^T$  being a vector of  $3 \times 1$ . It is easy to check that

$$\widetilde{Q}(\omega)^*\widetilde{Q}(\omega) = I_3 - \widetilde{P}(\omega)\widetilde{P}(\omega)^*$$

by using (28). If we choose the top  $2 \times 2$  principal block from the right-hand side of the above equation, we will have, letting  $\widetilde{Q} = [\widetilde{Q}_{jk}]_{1 \leq j,k \leq 3}$ ,

$$\begin{aligned} [\widetilde{Q}_{jk}]_{1 \leq j \leq 3}^* [\widetilde{Q}_{jk}]_{1 \leq j \leq 3} &= I_2 - \begin{bmatrix} \sqrt{2}P_1(\omega) \\ \sqrt{2}P_2(\omega) \end{bmatrix} \\ &\quad \begin{bmatrix} \sqrt{2}P_1(\omega)^* & \sqrt{2}P_2(\omega)^* \end{bmatrix}. \end{aligned}$$

Replacing  $\omega$  by  $2\omega$  in the above equation, multiplying the following unitary matrix from the left of the equation

$$U := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & z \\ 1 & -z \end{bmatrix}$$

and multiplying  $U^*$  from the right of the equation, we obtain

$$U [\widetilde{Q}_{jk}]_{1 \leq j \leq 3}^* [\widetilde{Q}_{jk}]_{1 \leq j \leq 3} U^* = I_2 - \begin{bmatrix} P(\omega) \\ P(\omega + \pi) \end{bmatrix} \begin{bmatrix} P(\omega)^* & P(\omega + \pi)^* \end{bmatrix}.$$

Then  $Q_{\ell}(\omega) = \sqrt{2}(\widetilde{Q}_{\ell,1}(2\omega) + z\widetilde{Q}_{\ell,2}(2\omega))/2$  for  $\ell = 1, 2, 3$  together with  $P(\omega)$  satisfy the matrix format UEP (cf. Lemma 8), and thus  $\psi_{\ell}$  defined by using these  $Q_{\ell}$  are tight wavelet frame generators.

Another method to construct tight wavelet frames is to do matrix extension. That is, let  $\widetilde{P}$  be the vector of size  $3 \times 1$  defined above. Note that  $\widetilde{P}^*\widetilde{P} = 1$ . By a method in [62], one can find  $\widetilde{Q}_1 = [\widetilde{Q}_{11}, \widetilde{Q}_{12}, \widetilde{Q}_{13}]^T$  and  $\widetilde{Q}_2 = [\widetilde{Q}_{21}, \widetilde{Q}_{22}, \widetilde{Q}_{23}]^T$ , two vectors of size  $3 \times 1$  such that the matrix  $[\widetilde{P} \ \widetilde{Q}_1 \ \widetilde{Q}_2]$  is unitary. See also [14] for detail construction. Once we have such unitary extension, we know that the first two rows are unitary. Rewriting the row vectors in column format, we have

$$\begin{bmatrix} \sqrt{2}P_1(\omega) & \sqrt{2}P_2(\omega) \\ \widetilde{Q}_{11}(\omega) & \widetilde{Q}_{12}(\omega) \\ \widetilde{Q}_{21}(\omega) & \widetilde{Q}_{22}(\omega) \end{bmatrix}^* \begin{bmatrix} \sqrt{2}P_1(\omega) & \sqrt{2}P_2(\omega) \\ \widetilde{Q}_{11}(\omega) & \widetilde{Q}_{12}(\omega) \\ \widetilde{Q}_{21}(\omega) & \widetilde{Q}_{22}(\omega) \end{bmatrix} = I_2.$$

Substituting  $\omega$  by  $2\omega$  in the above equation and multiplying  $\widetilde{U}$  from the right and  $\widetilde{U}^*$  from the left of the above equation both sides, we obtain

$$\begin{bmatrix} P(\omega) & P(\omega + \pi) \\ Q_1(\omega) & Q_1(\omega + \pi) \\ Q_2(\omega) & Q_2(\omega + \pi) \end{bmatrix}^* \begin{bmatrix} P(\omega) & P(\omega + \pi) \\ Q_1(\omega) & Q_1(\omega + \pi) \\ Q_2(\omega) & Q_2(\omega + \pi) \end{bmatrix} = I_2$$

which is just another form of the UEP (see (24)). Here  $Q_{\ell}(\omega) = \widetilde{Q}_{\ell,1}(2\omega) + z\widetilde{Q}_{\ell,2}(2\omega)$  for  $\ell = 1$  and  $\ell = 2$ . Hence, we can define two tight wavelet frame generators  $\psi_{\ell}$  using the above  $Q_{\ell}$  for  $\ell = 1, 2$ .

We now discuss a new method how to construct tight wavelet frames. Let

$$\widehat{Q}(\omega) = I_2 - \widehat{P}(\omega)\widehat{P}(\omega)^*$$

where  $\widehat{P}(\omega) = [\sqrt{2}P_1(\omega), \sqrt{2}P_2(\omega)]^T$ . We claim that  $\widehat{Q}$  is nonnegative definite under the assumption (26). Clearly,  $\widehat{Q}(\omega)$  is symmetric. 1 is an eigenvalue of  $\widehat{Q}$  since any vector  $v(\omega)$  orthogonal to  $\widehat{P}(\omega)$  is an eigenvector of  $\widehat{Q}$ .  $1 - \widehat{P}(\omega)^*\widehat{P}(\omega) \geq 0$  is another eigenvalue of  $\widehat{Q}$ . Thus, by matrix-valued Riesz–Fejér factorization (cf. [34] and [30]) we have

$$\widehat{Q}(\omega) = \widetilde{Q}(\omega)\widetilde{Q}(\omega)^*$$

for a Laurent polynomial matrix  $\widetilde{Q}(\omega)$  of size  $2 \times 2$ . That is,

$$\widehat{P}(\omega)\widehat{P}(\omega)^* + \widetilde{Q}\widetilde{Q}^* = I_2.$$

Replacing  $\omega$  by  $2\omega$  in the above equation, multiplying  $U$  from the left and  $U^*$  from the right of the equation we have

$$\begin{aligned} & \begin{bmatrix} P(\omega) \\ P(\omega + \pi) \end{bmatrix} [P(\omega)^* P(\omega + \pi)^*] \\ & + \begin{bmatrix} Q_1(\omega) & Q_1(\omega + \pi) \\ Q_2(\omega) & Q_2(\omega + \pi) \end{bmatrix} \\ & \times \begin{bmatrix} Q_1(\omega)^* & Q_2(\omega) \\ Q_1(\omega + \pi)^* & Q_2(\omega + \pi) \end{bmatrix} = I_2. \end{aligned}$$

We now use B-splines to give some examples of tight wavelet frames (cf. [14]).

**Example 16** Consider linear B-splines. The mask polynomial  $P(\omega) = (1 + z)^2/4$  with  $z = e^{i\omega}$ . It is easy to find  $Q_1(z) = -1/4 + z/2 - z^2/4$  and  $Q_2(z) = \sqrt{2}(1 - z^2)/2$ . One of the tight wavelet frame generators defined by using  $Q_1$  and  $Q_2$  is symmetric and the other is anti-symmetric.

**Example 17** Consider quadratic B-splines. The mask polynomial  $P(\omega) = (1 + z)^3/8$ . One can find  $Q_1(z) = \sqrt{3}(1 - z)/4$  and  $Q_2(z) = (1 + 3z - 3z^2 - z^3)/8$ . The tight wavelet generators associated with these  $Q_1$  and  $Q_2$  are all anti-symmetric.

**Example 18** Consider cubic B-splines. The mask polynomial  $P(\omega) = (1 + z)^4/16$ . Denote

$$a = \frac{\sqrt{8 - 2\sqrt{14}}}{8}, b = \frac{\sqrt{8 + 2\sqrt{14}}}{8}, r = \frac{\sqrt{16 + 2\sqrt{14}}}{8}.$$

Let

$$\begin{aligned} Q_1(z) &= 4ar^2 + \left(r - \frac{1}{16r}\right) \frac{z}{\sqrt{2}} + \frac{1 - 2r^2}{\sqrt{2}r} z^2 \\ &\quad - \frac{b}{\sqrt{2}} z^3 - \frac{b}{4\sqrt{2}} z^4 \\ Q_2(z) &= \frac{r}{4\sqrt{2}} + \left(a + \frac{b}{\sqrt{2}}\right) z \\ &\quad - \frac{b}{2\sqrt{2}} z^2 - \frac{b}{\sqrt{2}r} z^3 - \frac{b^2}{\sqrt{2}r} z^4. \end{aligned}$$

One can define tight wavelet frame generators  $\psi_\ell$  using these  $Q_\ell$ .

It is also possible to find symmetric and anti-symmetric tight wavelet frame generators using B-splines. We refer to [14,72,73] for the detail.

### Tight Wavelet Frames over Bounded Domain

We discuss the construction of tight wavelet frames over bounded domain. The following discussion works for a bounded domain over any Euclidean space. For convenience, we restrict our attention to the univariate setting.

Let  $\Omega \subset \mathbb{R}$  be a bounded domain, e.g., an interval. A tight wavelet frame for  $L_2(\Omega)$  is based on a half infinite sequence of nested subspaces over  $\Omega$ . Suppose that we have a sequence of nested subspaces  $\{V_k\}_{k \in \mathbb{Z}_+} \subset L_2(\Omega)$  satisfying

$$\begin{aligned} V_1 \subset V_2 \subset \dots \subset V_k \subset \dots \rightarrow L_2(\Omega) \quad \text{and} \\ \bigcup_{k=1}^{\infty} V_k \text{ is dense in } L_2(\Omega). \end{aligned}$$

Let  $\Phi_k := (\phi_{k,1}, \dots, \phi_{k,m_k})^T$  be a column vector of locally supported functions in  $V_k$  which generate  $V_k$ , i.e.,  $V_k = \text{span}\{\phi_{k,1}, \dots, \phi_{k,m_k}\}$ .

Because  $V_k$  is a subspace of  $V_{k+1}$ , the vector  $\Phi_k$  in  $V_k$  can be generated by the column vector  $\Phi_{k+1}$  which spans  $V_{k+1}$ . Thus there exists a matrix  $P_k$  of size  $m_k \times m_{k+1}$  with  $m_k \leq m_{k+1}$  such that

$$\Phi_k = P_k \Phi_{k+1}. \quad (29)$$

The matrix  $P_k$  is often called a refinement matrix. Let  $Q_k$  be another matrix of size  $n_k \times m_{k+1}$ . Define

$$\Psi_k := Q_k \Phi_{k+1}. \quad (30)$$

**Definition 3** The family of vectors  $\{\Psi_k\}_{k \in \mathbb{Z}_+}$  defined in (30) is a *tight wavelet frame* associated with  $\{\Phi_k\}_{k \in \mathbb{Z}_+}$  in

$L_2(\Omega)$  if

$$\|f\|^2 = \sum_{j=1}^{m_1} |\langle f, \phi_{1,j} \rangle|^2 + \sum_{k=1}^{\infty} \sum_{j=1}^{n_k} |\langle f, \psi_{k,j} \rangle|^2, \\ \forall f \in L_2(\Omega),$$

where  $\langle f, g \rangle = \int_{\Omega} f(x)g(x)dx$  be the standard inner product on  $L_2(\Omega)$ .

If  $\phi_{1,j}, j = 1, \dots, m_1$  and  $\psi_{k,j}, j = 1, \dots, n_k, k = 1, \dots$  generate a tight wavelet frame, then one can prove that

$$f = \sum_{j=1}^{m_1} \langle f, \phi_{1,j} \rangle \phi_{1,j} + \sum_{k=1}^{\infty} \sum_{j=1}^{n_k} \langle f, \psi_{k,j} \rangle \psi_{k,j}$$

for any  $f \in L_2(\Omega)$  by using the polarization technique mentioned before.

We show that a matrix  $Q_k$  satisfying

$$I_{m_{k+1}} = P_k^T P_k + Q_k^T Q_k, \quad (31)$$

for a given refinement matrix  $P_k$  in (29) is a key step for constructing a tight wavelet frame. Here  $I_{m_{k+1}}$  is the standard identity matrix of size  $m_{k+1}$ .

Clearly, each component in the function vector  $\Psi_k$  is in  $V_{k+1}$ . We want to have

$$\langle f, \Phi_{k+1} \rangle^T \Phi_{k+1} = \langle f, \Phi_k \rangle^T \Phi_k + \langle f, \Psi_k \rangle^T \Psi_k, \\ \forall f \in L_2(\Omega). \quad (32)$$

Let  $c_{k,i} := \langle f, \phi_{k,i} \rangle$  for all  $i = 1, \dots, m_k$  and  $C_k := (c_{k,1}, \dots, c_{k,m_k})^T$  be a column vector of size  $m_k \times 1$  for any  $k \in \mathbb{Z}_+$ . In the same way, let  $d_{k,j} := \langle f, \psi_{k,j} \rangle$  for all  $j = 1, \dots, n_k$  and  $D_k := (d_{k,1}, \dots, d_{k,n_k})^T$ . Then we know

$$C_k = \langle f, \Phi_k \rangle = \langle f, P_k \Phi_{k+1} \rangle = P_k C_{k+1}, \quad (33)$$

$$D_k = \langle f, \Psi_k \rangle = \langle f, Q_k \Phi_{k+1} \rangle = Q_k C_{k+1}. \quad (34)$$

Thus condition in (32) can be expressed in the following form

$$C_{k+1}^T \Phi_{k+1} = C_k^T P_k \Phi_{k+1} + D_k^T Q_k \Phi_{k+1}.$$

That is,  $C_{k+1}^T = C_k^T P_k + D_k^T Q_k$ . By using (34), we get

$$C_{k+1}^T C_{k+1} = C_{k+1}^T (P_k^T P_k + Q_k^T Q_k) C_{k+1}.$$

This implies that  $Q_k$  must satisfy (31) for all  $k \geq 1$ . On the other hand, if we find  $Q_k$  satisfying (31) for all  $k \geq 1$ , then we have the above equation and hence, by using (34),

$$C_{k+1}^T C_{k+1} = C_k^T C_k + D_k^T D_k.$$

It follows for any  $\ell \in \mathbb{Z}_+$  with  $\ell < k$ ,

$$C_{k+1}^T C_{k+1} = C_{\ell}^T C_{\ell} + \sum_{j=\ell}^k D_j^T D_j. \quad (35)$$

The condition (31) implies  $C_{k+1}^T = C_{k+1}^T (P_k^T P_k + Q_k^T Q_k) = C_k^T P_k + D_k^T Q_k$  and hence,

$$C_{k+1}^T \Phi_{k+1} = C_k^T \Phi_k + Q_k^T \Psi_k = \dots = C_{\ell}^T \Phi_{\ell} + \sum_{j=\ell}^k D_j^T \Psi_j.$$

If  $C_{k+1}^T \Phi_{k+1}$  converges to  $f$  in  $L_2(\Omega)$ , for any  $\ell \in \mathbb{Z}_+$ , we have

$$\|f\|^2 = \left\langle f, \lim_{k \rightarrow +\infty} C_{k+1}^T \Phi_{k+1} \right\rangle \\ = \lim_{k \rightarrow +\infty} \left\langle f, C_{\ell}^T \Phi_{\ell} + \sum_{j=\ell}^k D_j^T \Psi_j \right\rangle \\ = C_{\ell}^T C_{\ell} + \sum_{j=\ell}^{\infty} D_j^T D_j \\ = \sum_{j=1}^{m_{\ell}} |\langle f, \phi_{\ell,j} \rangle|^2 + \sum_{k=\ell}^{\infty} \sum_{j=1}^{n_k} |\langle f, \psi_{k,j} \rangle|^2.$$

**Theorem 13** Suppose that  $\Phi_k$  is a given refinable vector which spans  $V_k$  for all  $k \geq 1$  with refinable matrix  $P_k$ , i. e.,  $\Phi_k = P_k \Phi_{k+1}$ . Suppose that the projections  $f$  to  $V_k$  converge to  $f$ , i. e.,  $C_k^T \Phi_k \rightarrow f$  for any  $f \in L_2(\Omega)$ . Suppose  $Q_k$  satisfies (31). Let  $\Psi_k = Q_k \Phi_k$ . Then  $\Psi_k, k \in \mathbb{Z}_+$  form a tight wavelet frame. Hence, any  $f \in L_2(\Omega)$  can be generated by using  $\Phi_{\ell}$  and  $\Psi_k$  with  $k \geq \ell$  for any  $\ell \geq 1$  using the formula above.

We now explain how to compute  $Q_k$  satisfying (31).

**Theorem 14** Let  $\{V_k\}_{k=1}^{\infty}$  be a nested sequence and  $V_k$  be generated by a family of functions  $\Phi_k$ . Denote by  $P_k$  the refinable matrix, i. e.,  $\Phi_k = P_k \Phi_{k+1}$ . If  $I_{m_k} - P_k P_k^T$  is positive semi-definite, then there exists a matrix  $Q_k$  satisfying (31). Moreover, if each component function  $\phi_{k,j}$  of  $\Phi_k$  is locally supported then each component function  $\psi_{k,j}$  of the vector  $\Psi_k$  is locally supported.

We note that the condition  $I_{m_k} - P_k P_k^T \geq 0$  is different from (31) which may be rewritten in  $I_{m_{k+1}} - P_k^T P_k = Q_k^T Q_k \geq 0$  since the size of the matrix, e. g.,  $I_{m_k}$  is smaller than that of  $I_{m_{k+1}}$ . The proof of Theorem 14 can be found in [54].

Next we use B-splines to construct tight wavelet frames over a bounded interval  $[0, b]$  for an integer  $b > 0$ . Fix



$m \geq 1$  and consider B-spline  $N_m$  which satisfies the following refinement equation:

$$N_m(x) = \sum_{j \in \mathbb{Z}} c_j^m N_m(2x - j),$$

where

$$c_j^m = \begin{cases} 2^{-m+1} \binom{m}{j}, & \text{for } 0 \leq j \leq m \\ 0 & \text{otherwise.} \end{cases} \quad (36)$$

We now define scaling functions  $\phi_{j,k}(x) = 2^{j/2} N_m(2^j x - k)|_{[0,b]}$  which are nonzero when  $k = 1 - m, \dots, 2^j b - m + 1$ . Thus, let

$$V_j := \{\phi_{j,k} : -1 \leq k \leq 2^j b - m + 1\}$$

for all  $j$ . It is easy to see that  $V_j \subset V_{j+1}$  by using (36). Also, we can easily verify that  $\bigcup_{j=1}^{\infty} V_j$  is dense in  $L_2([0, b])$ . Let  $m_j = 2^j b - m + 2$  and  $\Phi_j = [\phi_{j,-1}, \dots, \phi_{j,m_j-1}]^T$  be a vector of basis functions for  $V_j$ . We can find a refinement matrix  $P_j^m$  of size  $m_j \times m_{j+1}$  satisfying  $\Phi_j^m = P_j^m \Phi_{j+1}^m$  for each  $j \in \mathbb{Z}_+$  using the coefficients in (36). Based on the discussion above we can check the positive semi-definite property of the matrix

$$I_{m_j} - P_j^m \cdot P_j^{mT}$$

for the identity matrix  $I_{m_j}$ .

**Lemma 9** The above matrix of size  $m_j \times m_j$  associated with B-spline  $N_m$  of order  $m$  is positive semi-definite for each  $j \in \mathbb{Z}$  and  $m \geq 1$ .

*Proof* Let us denote  $(p_{i,k}^{m,j}) := P_j^m$ . Then for each  $i = 1, \dots, m_j$

$$0 \leq \sum_{k=1}^{m_{j+1}} p_{i,k}^{m,j} \leq \frac{1}{2} \sum_{k=0}^m c_k^m = 1, \quad (37)$$

where  $c_k^m$  is in (36). Let us express  $[g_{i,k}^{m,j}]_{1 \leq i, k \leq m_j} = P_j^m \cdot P_j^{mT}$ . We claim that the matrix  $I_{m_j} - P_j^m \cdot P_j^{mT}$  is diagonally dominant. Indeed, it is sufficient to check  $|1 - g_{i,i}^{m,j}| \geq \sum_{k \neq i} |g_{i,k}^{m,j}|$  for  $i \leq m_k$ . Notice that

$$g_{i,k}^{m,j} = \sum_{\ell=1}^{m_{j+1}} p_{i,\ell}^{m,j} p_{\ell,k}^{m,j}.$$

Then for each  $k \in \mathbb{Z}_+$ ,

$$\begin{aligned} 1 - |g_{i,i}^{m,j}| - \sum_{k \neq i} |g_{i,k}^{m,j}| &= 1 - \sum_{k=1}^{m_{j+1}} \sum_{\ell=1}^{m_{j+1}} p_{i,\ell}^{m,k} p_{\ell,k}^{m,j} \\ &= 1 - \left( \sum_{\ell=1}^{m_{j+1}} p_{i,\ell}^{m,j} \right)^2. \end{aligned}$$

Since (37),  $1 - |g_{i,i}^{m,j}| \geq \sum_{k \neq i} |g_{i,k}^{m,j}|$  for all  $i = 1, \dots, m_k$ . Therefore the symmetry matrix  $I_{m_k} - P_k^m \cdot P_k^{mT}$  is positive semi-definite. This completes the proof.  $\square$

That is, we can use B-splines of any order to construct tight wavelet frames over any intervals.

### q-Dilated Orthonormal Wavelets

In the previous sections, we consider refinable functions  $\phi$  which are dilated by 2. That is,

$$\phi(x) = \sum_n h_n \sqrt{q} \phi(qx - n), \quad x \in \mathbb{R} \quad (38)$$

for  $q = 2$ . In general, we can consider refinable function  $\phi$  which can be dilated by  $q > 2$ . All the previous theories of various wavelets have a generalization in this setting of dilation factor  $q > 2$ . One important feature of wavelets in the new setting is that it requires more than one wavelet function. Let us give a brief explanation of orthonormal wavelets based on dilation factor  $q = 3$ .

Suppose we have a multiresolution approximation of  $L_2(\mathbb{R})$  based on dilation factor  $q$ . That is, we have

**Definition 4** A sequence of closed subspaces  $\dots V_{-2}, V_{-1}, V_0, V_1, V_2, \dots$  of  $L_2(\mathbb{R})$  is a multiresolution approximation of  $L_2(\mathbb{R})$  if they satisfy the following

1.  $V_j \subset V_{j+1}, \forall j \in \mathbb{Z}$ ;
2.  $\bigcup_{j \in \mathbb{Z}} V_j = L_2(\mathbb{R})$  and  $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$ ;
3.  $f(\cdot) \in V_0 \iff f(q^j \cdot -k) \in V_j$  for all  $j, k \in \mathbb{Z}$ ;
4. There is a function  $\phi \in L_2(\mathbb{R})$  called scaling function such that  $\{\phi(x - k), k \in \mathbb{Z}\}$  forms an orthonormal basis of  $V_0$ .

By taking the Fourier transform both sides of (38), we get that

$$\hat{\phi}(\xi) = m_0(\omega/q) \hat{\phi}(\omega/q), \quad \xi \in \mathbb{R}, \quad (39)$$

where

$$m_0(\omega) = \frac{1}{\sqrt{q}} \sum_n h_n e^{-in\omega}. \quad (40)$$

$m_0$  is a Laurent polynomial in  $z = e^{i\omega}$  and is called the mask of  $\phi$ .

Since  $\phi$  is orthonormal, we know

$$\begin{aligned} \delta_{n0} &= \int_{\mathbb{R}} \phi(x) \overline{\phi(x - n)} dx = \frac{1}{2\pi} \int_{\mathbb{R}} |\hat{\phi}(\xi)|^2 e^{in\xi} d\xi \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left( \sum_{k \in \mathbb{Z}} |\hat{\phi}(\xi + 2k\pi)|^2 \right) e^{in\xi} d\xi, \end{aligned}$$

and hence,

$$\sum_{k \in \mathbb{Z}} |\hat{\phi}(\xi + 2k\pi)|^2 = 1. \quad (41)$$

Since  $m_0$  is  $2\pi$ -periodic, by (39) and (41) we find that

$$\begin{aligned} 1 &= \sum_{k \in \mathbb{Z}} |\hat{\phi}(q\xi + 2k\pi)|^2 \\ &= \sum_{k \in \mathbb{Z}} |m_0(\xi + 2k\pi/q)|^2 |\hat{\phi}(\xi + 2k\pi/q)|^2 \\ &= |m_0(\xi)|^2 \sum_{k \in \mathbb{Z}} |\hat{\phi}(\xi + 2k\pi)|^2 + |m_0(\xi + 2\pi/q)|^2 \\ &\quad \sum_{k \in \mathbb{Z}} |\hat{\phi}(\xi + 2\pi/q + 2k\pi)|^2 \\ &\quad + \dots + |m_0(\xi + 2(q-1)\pi/q)|^2 \\ &\quad \sum_{k \in \mathbb{Z}} |\hat{\phi}(\xi + 2(q-1)\pi/q + 2k\pi)|^2. \end{aligned}$$

Hence

$$|m_0(\xi)|^2 + |m_0(\xi + 2\pi/q)|^2 + \dots + |m_0(\xi + 2(q-1)\pi/q)|^2 = 1. \quad (42)$$

Consider  $j = 0$  and let  $W_0$  be the orthogonal complement of  $V_0$  in  $V_1$ . Since  $V_0$  is generated by integer translates of  $\phi(x)$  and  $V_1$  is generated by integer translates of  $q$  functions,  $\phi(qx)$ ,  $\phi(qx-1)$ ,  $\dots$ ,  $\phi(qx-q+1)$ , we need  $q-1$  functions whose integer translates to generate  $W_0$ . That is, we look for  $\psi_1, \psi_2, \dots, \psi_{q-1}$  such that

$$\{\psi_i(x-k) : i = 1, 2, \dots, q-1, k \in \mathbb{Z}\} \quad (43)$$

forms an orthonormal basis for  $W_0$ . For convenience, let

$$W_0^i := \text{span}\{\psi_i(x-k), k \in \mathbb{Z}\} \quad \text{for } i = 1, 2, \dots, q-1. \quad (44)$$

Then we have  $W_0 = W_0^1 \oplus W_0^2 \oplus \dots \oplus W_0^{q-1}$ . In general, we set

$$W_j^i = \text{span}\{\psi_i(q^j x - k), k \in \mathbb{Z}\} \quad \text{for } i = 1, 2, \dots, q-1.$$

Then we have

$$V_{j+1} = V_j \oplus W_j^1 \oplus W_j^2 \oplus \dots \oplus W_j^{q-1} \quad (45)$$

and

$$W_j^i \perp W_j^{i'} \quad \text{if } i \neq i'. \quad (46)$$

It implies

$$L_2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} \bigoplus_{i=1}^{q-1} W_j^i. \quad (47)$$

That is,  $\psi_i, i = 1, 2, \dots, q-1$  form a set of orthonormal  $q$ -wavelet functions.

Since  $\psi_j \in V_1$  and  $\psi_j \perp V_0$ , we have

$$\psi_j(x) = \sqrt{q} \sum_n g_n^j \phi(qx - n), \quad (48)$$

with  $g_n^j = \langle \psi_j, \phi_{1,n} \rangle$  and  $\{g_n^j\} \in \ell^2(\mathbb{Z})$ . This implies

$$\hat{\psi}_j(\xi) = m_j(\xi/q) \hat{\phi}(\xi/q), \quad (49)$$

where

$$m_j(\xi) = \frac{1}{\sqrt{q}} \sum_n g_n^j e^{-in\xi}. \quad (50)$$

We can prove the following

**Lemma 10** Let  $\phi$  be an orthonormal refinable function associated with dilation factor  $q \geq 2$  and  $m_0$  be the corresponding mask. Let  $\psi_1, \psi_2, \dots, \psi_{q-1}$  be functions defined by in terms of Fourier transform using take  $m_1, m_2, \dots, m_{q-1}$  as above. Then the family  $\{\psi_j(x-n), j = 1, 2, \dots, q-1, n \in \mathbb{Z}\}$  is an orthonormal basis for the orthogonal complement  $W_0$  of  $V_0$  in  $V_1$  if and only if the matrix

$$\begin{bmatrix} m_0(\omega) & m_0(\omega + 2\pi/q) \\ m_1(\omega) & m_1(\omega + 2\pi/q) \\ \vdots & \vdots \\ m_{q-1}(\omega) & m_{q-1}(\omega + 2\pi/q) \\ \dots & m_0(\omega + 2(q-1)\pi/q) \\ \dots & m_1(\omega + 2(q-1)\pi/q) \\ \ddots & \vdots \\ \dots & m_{q-1}(\omega + 2(q-1)\pi/q) \end{bmatrix} \quad (51)$$

is unitary.

As usual, the first problem is to find a Laurent polynomial  $m_0$  such that  $\sum_{i=0}^{q-1} |m_0(\omega + 2i\pi/q)|^2 = 1$ . Once such  $m_0$  is found, we can find the other  $m_j, j = 1, \dots, q-1$  by using unitary matrix extension technique in [62].

To give some examples, we now restrict ourselves to  $q = 3$ . Consider a mask  $m_0$  which must satisfy

$$m_0(0) = 1$$

in order to make  $\prod_{j=1}^{\infty} m_0(\omega/3^j)$  converges so that it defines a distribution  $\phi$ .  $m_0$  must satisfy

$$|m_0(\omega)|^2 + |m_0(\omega + 2\pi/3)|^2 + |m_0(\omega + 4\pi/3)|^2 = 1$$

so that  $\phi$  is in  $L_2(\mathbb{R})$ . Note that the above two conditions implies that  $m_0(\omega) = \left(\frac{1+e^{i\omega}+e^{2i\omega}}{3}\right)\mathcal{L}(\omega)$  for some Laurent polynomial in  $e^{i\omega}$ . The following can be found in [92].

**Lemma 11** A Laurent polynomial  $m_0(\omega) = \sum_{k=0}^5 p_k e^{ik\omega}$  satisfies above two conditions if and only if

$$\begin{aligned} p_0 &= \frac{1}{6} + \frac{\sqrt{3}}{6} \cos \theta, & p_1 &= \frac{1}{6} + \frac{\sqrt{3}}{6} \sin \theta \cos \alpha \\ p_2 &= \frac{1}{6} + \frac{\sqrt{3}}{6} \sin \theta \sin \alpha, & p_3 &= \frac{1}{6} - \frac{\sqrt{3}}{6} \cos \theta \\ p_4 &= \frac{1}{6} - \frac{\sqrt{3}}{6} \sin \theta \cos \alpha, & p_5 &= \frac{1}{6} - \frac{\sqrt{3}}{6} \sin \theta \sin \alpha \end{aligned}$$

for some  $\theta, \alpha$  in  $[0, 2\pi]$ .

**Example 19** A scaling function associated with  $m_0$  in Lemma 11 is symmetric if and only if the coefficients given above with  $\tan \theta = -1$ ,  $\sin \alpha = 1$  or  $\tan \theta = 1$ ,  $\sin \alpha = -1$ .

**Example 20**  $m_0$  in Lemma 11 has two order of vanishing moments, i.e.,  $m_0$  contains a factor  $(1 + e^{i\omega} + e^{i2\omega})^2$  if  $\alpha = \arccos\left(\sqrt{\frac{19}{50+4\sqrt{57}}}\right)$  and  $\theta = \arcsin\left(\frac{\sqrt{50+4\sqrt{57}}}{9}\right)$ .

Next we consider  $m_0(z) = p_0 + p_1 z + p_2 z^2 + p_3 z^3 + p_4 z^4 + p_5 z^5 + p_6 z^6 + p_7 z^7 + p_8 z^8$  be the symbol of a scaling function supported on  $[0, 4]$ , where  $z = e^{i\omega}$ .

**Lemma 12**  $m_0(z)$  satisfies  $m_0(0) = 1$  and

$$|m_0(z)|^2 + |m_0(z\tau)|^2 + |m_0(z\tau^2)|^2 = 1, \quad \forall z = e^{i\omega}, \quad \omega \in \mathbb{R}, \quad (52)$$

with  $\tau = e^{i\frac{2\pi}{3}}$  if and only if

$$\begin{aligned} p_1 &= \frac{1}{12} + \frac{\sqrt{3}}{12} \sin \alpha + \frac{1}{6} r \sin \theta \\ p_2 &= \frac{1}{12} + \frac{\sqrt{3}}{12} \cos \alpha \sin \beta + \frac{1}{6} r \cos \theta \sin \gamma \\ p_3 &= \frac{1}{6} (1 - \sqrt{3} \cos \alpha \cos \beta) \\ p_4 &= \frac{1}{6} (1 - \sqrt{3} \sin \alpha) \\ p_5 &= \frac{1}{6} (1 - \sqrt{3} \cos \alpha \sin \beta) \\ p_6 &= \frac{1}{12} + \frac{\sqrt{3}}{12} \cos \alpha \cos \beta - \frac{1}{6} r \cos \theta \cos \gamma \\ p_7 &= \frac{1}{12} + \frac{\sqrt{3}}{12} \sin \alpha - \frac{1}{6} r \sin \theta \\ p_8 &= \frac{1}{12} + \frac{\sqrt{3}}{12} \cos \alpha \sin \beta - \frac{1}{6} r \cos \theta \sin \gamma, \end{aligned}$$

where

$$r = \sqrt{\frac{1}{2} + \frac{\sqrt{3}}{6} (\cos \alpha \cos \beta + \cos \alpha \sin \beta + \sin \alpha)} \quad (53)$$

for some  $\alpha, \beta, \gamma$  in  $[0, 2\pi]$ .

**Example 21** A scaling function associated with  $m_0$  in Lemma 12 is symmetric if and only if the coefficients given above with  $\sin \theta = 0$ ,  $\cos \alpha = 0$ ,  $\tan \beta = 1$ , and  $\tan \gamma = -1$ .

**Example 22**  $m_0$  in Lemma 12 has three order of vanishing moments, i.e.,  $m_0$  contains a factor  $(1 + e^{i\omega} + e^{i2\omega})^3$  if  $\alpha = \gamma = 0$ ,  $\beta = \arcsin(\frac{7\sqrt{6}}{36}) - \frac{\pi}{4}$ , and  $\theta = \pi + \arcsin(\frac{3}{5})$ .

### Multiwavelets and Balanced Multiwavelets

We shall explain a construction of multiwavelets in this section. Typical multiwavelets are DGHM multiwavelets (cf. [28]), Chui-Lian wavelets (cf. [16]) and multiwavelets based on B-splines (cf. [29]). In the following we present a newer construction. An advantage is that the number of wavelets is always 3 no matter how smooth the wavelets are.

Fix integer  $r > 1$ . Let  $\Phi = [\phi_1, \dots, \phi_r]^T$  be a vector of compactly supported real-valued functions in  $\mathbb{R}$ . We suppose that  $\Phi$  is refinable. That is, there exist matrices  $A_k$ 's of size  $r \times r$  such that

$$\Phi(x) = \sum_{k \in \mathbb{Z}^d} A_k \Phi(2x - k), \quad x \in \mathbb{R}.$$

Also, we say  $\Phi$  is orthonormal if

$$\int_{\mathbb{R}} \phi_i(x) \phi_j(x - k) dx = \begin{cases} 1, & \text{if } i = j \text{ and } k = 0, \\ 0, & \text{otherwise} \end{cases}$$

for all  $i, j = 1, \dots, r$ .  $\Phi$  generates a space  $S$  if  $S$  consists of all finitely linear combination of integer translates of entries of  $\Phi$ .

Next we define a Grammian matrix  $G = (G_{ij})_{i,j=1,\dots,r}$  of size  $r \times r$  associated with  $\Phi$  by

$$G_{ij}(z) = \sum_{k \in \mathbb{Z}} z^k \int_{\mathbb{R}} \phi_i(x) \phi_j(x - k) dx$$

for all  $i, j = 1, \dots, r$  with  $z \in \mathbb{C} \setminus \{0\}$ . We note that  $\Phi$  is orthonormal if and only if its Grammian matrix  $G$  is the identity of size  $r \times r$ .

We suppose that  $\Phi$  generates a space  $S$ . Then for any compactly supported functions  $\psi_1, \dots, \psi_s$  in  $S$ , there exists a finitely many nonzero matrices  $C_k$  of size  $s \times r$  such

that

$$\Psi(x) = [\psi_1(x), \dots, \psi_s(x)]^T = \sum_{k \in \mathbb{Z}} C_k \Phi(x - k).$$

In terms of Fourier transform, we have

$$\widehat{\Psi}(\omega) = C(z) \widehat{\Phi}(\omega)$$

where  $C(z)$  denotes the  $s \times r$  matrix of Laurent polynomials, i. e.,

$$C(z) := \sum_{k \in \mathbb{Z}} C_k z^k.$$

A square matrix  $C(z)$  is said to be invertible if  $\det(C(z))$  is a monomial of  $z$ , e. g.,  $\alpha z^m$  for a scalar  $\alpha \neq 0$  and an integer  $m \in \mathbb{Z}$ . It is clear that if  $C(z)$  is invertible,  $\Psi$  generates the same  $S$ . A proof of the following result can be found in literature (cf. e. g., [31]).

**Lemma 13** Suppose that  $\Psi$  is a vector of compactly supported functions and generates a space  $S$ . Let  $G(z) = (G_{ij}(z))_{i,j=1,\dots,r}$  of size  $r \times r$  by

$$G_{ij}(z) = \sum_{k \in \mathbb{Z}} z^k \int_{\mathbb{R}} \psi_i(x) \psi_j(x - k) dx$$

for all  $i, j = 1, \dots, r$  be the Grammian matrix associated with  $\Psi$ . If the determinant of the Grammian matrix  $G(z)$  is a nonzero constant, then there exists a  $\Phi$  which is orthonormal and generates  $S$ . The converse is also true.

The above lemma reveals a key for constructing orthonormal vector of scaling functions: find  $\psi_1, \dots, \psi_r$  which generate the same space  $S$  such that its Grammian matrix has a constant determinant.

We now explain how to use B-splines for constructing an orthonormal vector of scaling functions with  $r = 3$ . Let  $N_m$  be the uniform B-spline of order  $m$ , in terms of Fourier transform,

$$\widehat{N}_m(\omega) = \left( \frac{1 - e^{-i\omega}}{i\omega} \right)^m.$$

Let  $V_0 = \text{span}\{N_m(x - k), k \in \mathbb{Z}\}$  be the spline space. Since  $N_m$  is a refinable function, for  $V_1$  being spanned by the integer translates of  $N_m(2x - k)$ ,  $k \in \mathbb{Z}$ , we have  $V_0 \subset V_1$ . Thus, letting  $\psi_1(x) = N_m(2x)$  and  $\psi_2(x) = N_m(2x - 1)$ ,  $\psi_1$  and  $\psi_2$  generate  $V_1$ . On the other hand, by the dilation equation, there exist two finite sequences  $a_{2k}$  and  $a_{2k+1}$  such that

$$N_m(x) = \sum_{k \in \mathbb{Z}} a_{2k} \psi_1(x - k) + \sum_{k \in \mathbb{Z}} a_{2k+1} \psi_2(x - k).$$

Note that the Fourier transform of the above equation is

$$\widehat{N}_m(2\omega) = \frac{1}{2} A(z) \widehat{N}_m(\omega)$$

and

$$\begin{aligned} \widehat{N}_m(\omega) &= A_0(z) \widehat{\psi}_1(\omega) + A_1(z) \widehat{\psi}_2(\omega) \\ &= A_0(z) \frac{1}{2} \widehat{N}_m\left(\frac{\omega}{2}\right) + A_1(z) \frac{1}{2} z^{\frac{1}{2}} \widehat{N}_m\left(\frac{\omega}{2}\right) \end{aligned}$$

where

$$A_0(z) = \sum_{k \in \mathbb{Z}} a_{2k} z^k \quad \text{and} \quad A_1(z) = \sum_{k \in \mathbb{Z}} a_{2k+1} z^k.$$

It follows that

$$A(z) = A_0(z^2) + z A_1(z^2).$$

It is known that  $A(z) = 2 \left( \frac{1+z}{2} \right)^m$ . It is easy to see that there exist two Laurent polynomials  $B_0(z)$  and  $B_1(z)$  of degree  $\leq m$  such that

$$A_0(z) B_0(z) + A_1(z) B_1(z) = 1.$$

We now define a new spline function in terms of Fourier transform by

$$\widehat{M}_m(\omega) = -B_1(z) \widehat{\psi}_1(\omega) + B_0(z) \widehat{\psi}_2(\omega).$$

Recall that

$$\widehat{N}_m(\omega) = A_0(z) \widehat{\psi}_1(\omega) + A_1(z) \widehat{\psi}_2(\omega).$$

It follows that  $N_m$  and  $M_m$  generate  $V_1$  since the determinant of the following matrix

$$\begin{bmatrix} \widehat{N}_m(\omega) \\ \widehat{M}_m(\omega) \end{bmatrix} = \begin{bmatrix} A_0(z) & A_1(z) \\ -B_1(z) & B_0(z) \end{bmatrix} \begin{bmatrix} \widehat{\psi}_1(\omega) \\ \widehat{\psi}_2(\omega) \end{bmatrix}$$

is constant 1. Furthermore,  $N_m(2x)$ ,  $N_m(2x - 1)$ ,  $M_m(2x)$ ,  $M_m(2x - 1)$  generate  $V_2$ .

Define  $\psi_3 = \sum_{k \in \mathbb{Z}} \alpha_k M_m(2x - k)$  for some finitely many nonzero coefficients  $\alpha_k$ . We will show how to find such  $\alpha_k$  that the Grammian matrix associated with  $\{\psi_1, \psi_2, \psi_3\}$ ,

$$G(z) = \left( \sum_{k \in \mathbb{Z}} z^k \int_{\mathbb{R}} \psi_i(x) \psi_j(x - k) dx \right)_{i,j=1,2,3}$$

has a constant determinant. Let

$$r(z) = \sum_{k \in \mathbb{Z}} \alpha_k z^k.$$

The computation in [31] shows

$$4 \det G(z^2) = D(z)r(z)r(1/z) + D(-z)r(-z)r(-1/z) \quad (54)$$

where

$$D(z) = \frac{1}{2}(a(z^2) - zb(z^2))(a(z)^2 - zb(z)^2)$$

with

$$a(z) = \sum_{k \in \mathbb{Z}} z^k \int N_m(2x)N_m(2x - k)dx,$$

$$b(z) = \sum_{k \in \mathbb{Z}} z^k \int N_m(2x)N_m(2x - 2k - 1)dx.$$

Now we claim that there exists a polynomial  $p(z) \geq 0$  such that

$$D(z)p(z) + D(-z)p(-z) = 1.$$

Once we have such a  $p(z)$ , it follows from the Riesz-Féjer lemma that there exists a polynomial  $r(z)$  such that  $r(z)r(1/z) = p(z)$ . This  $r(z)$  is the polynomial we look for such that the determinant (54) of Grammian matrix  $G(z)$  is a nonzero constant.

The existence of  $p(z)$  satisfying the above properties is guaranteed by the following lemma (see a constructive proof from [32]).

**Lemma 14** *Let  $q$  be a polynomial of degree  $n$  with all its zeros in  $[1, \infty)$  having a positive leading coefficient. Then there exists a unique polynomial  $p$  with real coefficients of degree  $n - 1$  such that*

$$p(x)q(x) + p(1 - x)q(1 - x) = 1$$

for  $x \in [0, 1]$ . Moreover,  $(-1)^n p(x) > 0$  for  $x \in (0, 1)$ .

By letting  $x = 1 - (z + 1/z)/2$  we can convert  $D(z)$  into a polynomial in  $x$ . By studying the zeros of  $D(z)$  we can see that  $q(x) = D(z)$  satisfies the conditions in the above lemma (for details, see [31] and [9]). This shows that the Grammian matrix  $G(z)$  associated with  $\Psi = (\psi_1, \psi_2, \psi_3)^T$  is nonzero monomial. Hence, it can be factored into  $G(z) = B(z)^*B(z)$  with invertible polynomial matrix  $B(z)$  by matrix-valued Fejér-Riesz lemma (cf. [34]), where  $B(z)^*$  stands for the transpose and conjugate of  $B(z)$ . A straightforward computational method to do such factorizations can be found in [9] based on the theory developed in [30]. Letting

$$\widehat{\Phi}(z) = B(z)^{-1}\widehat{\Psi}(z),$$

we know that the Grammian matrix of  $\Phi$  is  $B(z)^{-1}G(z)B^{-1}(z)^*$  which is the identity matrix and hence  $\Phi = [\phi_1,$

$\phi_2, \phi_3]^T$  is an orthonormal refinable function vector. Next we explain how to compute the associated wavelets. We begin with

**Lemma 15**  *$\Phi$  is refinable. That is, letting*

$$\widetilde{\Phi}(x) = \sqrt{2}(\phi_1(2x), \phi_2(2x), \phi_3(2x), \\ \phi_1(2x - 1), \phi_2(2x - 1), \phi_3(2x - 1))^T,$$

there exists matrix coefficients  $p_i$  of size  $3 \times 6$  such that

$$\Phi(x) = \sum_{k \in \mathbb{Z}} p_k \widetilde{\Phi}(x - k) \quad \text{or} \quad \widehat{\Phi}(\omega) = P(z)\widehat{\widetilde{\Phi}}(\omega), \quad (55)$$

where  $P(z)$  is a matrix mask of size  $3 \times 6$ .

Since  $\Phi$  is of compact support, we may assume that only  $m + 1$  terms  $p_0, p_1, \dots, p_m$  are nonzero matrix coefficients. Then the orthogonal condition implies

$$0 = \int_{\mathbb{R}} \Phi(x)\Phi(x - k)^T dx \\ = \sum_{i,j=1}^m p_i \int_{\mathbb{R}} \widetilde{\Phi}(x - i)\widetilde{\Phi}^T(x - j - k) dx p_j^T \\ = \sum_{i,j=1}^m p_i \delta_{i,j+k} I_{6 \times 6} p_j^T = \sum_{i=k}^m p_i p_{i-k}^T,$$

for  $k = 1, \dots, m$ . In particular, we have

$$p_m p_0^T = 0. \quad (56)$$

We now use induction on  $m$  to explain how to construct three compactly supported orthonormal wavelets  $h_1, h_2, h_3 \in S_1$  such that

$$\mathcal{W} := \text{span}\{h_1(\cdot - i), h_2(\cdot - j), h_3(\cdot - k), i, j, k \in \mathbb{Z}\}$$

is the orthogonal complement of  $S$  in  $S_1$ . It is trivial when  $m = 0$ . Indeed, in this case,  $P(z) = p_0$  is a scalar matrix. We simply choose  $Q(z)$  to be a scalar matrix which is an orthonormal extension of  $p_0$ . Assume that for  $m \geq 1$ , when  $P_m(z) = \sum_{k=0}^m p_k z^k$  is an orthonormal matrix of  $3 \times 6$ , we can find  $Q_m(z)$  such that

$$\begin{bmatrix} P_m(z) \\ Q_m(z) \end{bmatrix}$$

is unitary. We now consider the case of  $m + 1$ :  $P_{m+1}(z) = \sum_{k=0}^{m+1} p_k z^k$  satisfying orthonormal properties. In particular, (56) implies that there exists a unitary matrix  $U_0$  of size  $6 \times 6$  such that  $p_0 U_0 = [0_{3 \times 3} \ \widehat{p}_0^b]$  and  $p_{m+1} U_0 = [\widehat{p}_{m+1}^a \ 0_{3 \times 3}]$ , where  $\widehat{p}_0^b$  is of size  $3 \times 3$  and the



same for  $\tilde{p}_{m+1}^a$ . Writing  $p_k U_0 = [\tilde{p}_k^a, \tilde{p}_k^b]$  with  $\tilde{p}_k^a$  and  $\tilde{p}_k^b$  being of size  $3 \times 3$ . Then

$$P_{m+1}(z)U_0 = \left[ \sum_{k=1}^{m+1} \tilde{p}_k^a z^k, \sum_{k=0}^m \tilde{p}_k^b z^k \right].$$

Let

$$U_1 := \begin{bmatrix} \frac{1}{z} I_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & I_{3 \times 3} \end{bmatrix}.$$

Then it follows that

$$\begin{aligned} P_{m+1}(z)U_0U_1 &= \left[ \sum_{k=0}^m \tilde{p}_{k+1}^a z^k, \sum_{k=0}^m \tilde{p}_k^b z^k \right] \\ &= \sum_{k=0}^m [\tilde{p}_{k+1}^a, \tilde{p}_k^b] z^k. \end{aligned}$$

That is,  $\tilde{P}_m(z) := P_{m+1}(z)U_0U_1$  has only  $m+1$  terms and is unitary. By induction, we can find an unitary extension  $\tilde{Q}_m(z)$  such that

$$\begin{bmatrix} P_{m+1}(z)U_0U_1 \\ \tilde{Q}_m(z) \end{bmatrix}$$

is unitary. Clearly,

$$\begin{bmatrix} P_{m+1}(z)U_0U_1 \\ \tilde{Q}_m(z) \end{bmatrix} U_1^* U_0^* = \begin{bmatrix} P_{m+1}(z) \\ \tilde{Q}_m(z)U_1^*U_0^* \end{bmatrix}$$

is also unitary. It follows that  $Q_{m+1}(z) := \tilde{Q}_m(z)U_1^*U_0^*$  is an unitary extension of  $P_{m+1}(z)$ . This completes the induction procedure. Several examples of multiwavelets and filters associated with these multiwavelets are given in [9].

Before we apply these multiwavelets, we need to balance them. The concept of balance was initially proposed in [63]. Mainly we need to make sure that a constant signal is reproduced by using balanced multiwavelets with constant coefficient vectors. The discussion and computation of balancing the above multiwavelets are presented in [38].

### Multivariate Orthonormal Wavelets

Univariate wavelets have found successful applications in signal processing. To apply wavelet methods to digital image processing, we have to construct bivariate wavelets. The most commonly used method is the tensor product of univariate wavelets. This construction leads to a separable wavelet which has a disadvantage of giving a particular importance to the horizontal and vertical directions. Much effort has been spent on constructing non-separable

bivariate wavelets in the last ten years. In this and the following three sections, we survey some methods for constructing bivariate and multivariate non-separable compactly supported orthonormal, biorthogonal, pre-wavelets as well as tight wavelet frames. All the discussion is based on the commonly used uniform dilation matrix  $2I_d$   $d$ -dimensional Euclidean space  $\mathbb{R}^d$ , where  $I_d$  is the identity of  $d \times d$  matrix. Due to the space limitation, we are not able to include all the constructions. We refer the reader to the following literature [21,33,45,46,87,88], for other methods of constructing nonseparable compactly supported orthonormal wavelets (cf. [48]).

All the construction of compactly supported orthonormal wavelets is based multiresolution analysis (MRA). For convenience, we restrict our attention to  $\mathbb{R}^2$ . To construct bivariate wavelets, we need to solve the following two mathematical problems.

(1) Find a Laurent polynomial

$$m_0(x, y) = \sum_{\substack{-M \leq j \leq M \\ -N \leq k \leq N}} c_{jk} x^j y^k$$

so normalized that  $m_0(1, 1) = 1$  satisfying

$$|m_0(x, y)|^2 + |m_0(-x, y)|^2 + |m_0(x, -y)|^2 + |m_0(-x, -y)|^2 = 1 \quad (57)$$

for  $x = e^{i\xi}$  and  $y = e^{i\eta}$  with  $\xi, \eta \in \mathbb{R}$ .

(2) Find another three Laurent polynomials  $m_j(x, y)$ ,  $j = 1, 2, 3$  such that the following matrix  $\mathcal{M}$  is unitary, i. e.,

$$\mathcal{M} = \begin{bmatrix} m_0(x, y) & m_0(-x, y) \\ m_1(x, y) & m_1(-x, y) \\ m_2(x, y) & m_2(-x, y) \\ m_3(x, y) & m_3(-x, y) \end{bmatrix} \quad (58)$$

is unitary.

The condition (58) is called the perfect reconstruction condition as in Sect. “Definition of Filters”.

Since  $m_0(1, 1) = 1$ , let

$$\hat{\phi}(\xi, \eta) = \prod_{k=1}^{\infty} m_0(e^{i\xi/2^k}, e^{i\eta/2^k}) \quad (59)$$

be the refinable function associated with  $m_0$ . Then the condition (57) implies that  $\phi \in L_2(\mathbb{R}^2)$ .

In order to see that  $\phi$  is orthonormal, we may apply the multidimensional generalization of Cohen's condition or Lawton's condition (cf. [24]). For simplicity, let us assume that  $\phi$  is orthonormal in the sense that

$$\int_{\mathbb{R}^2} \phi(x, y) \phi(x-k, y-j) dx dy = \begin{cases} 1, & \text{if } j = k = 0 \\ 0, & \text{otherwise} \end{cases}$$

for all  $j, k \in \mathbb{Z}$  and  $\phi$  generate a multiresolution approximation of  $L_2(\mathbb{R}^2)$ . That is, let

$$V_j := \text{span}\{2^j \phi(2^j x - m, 2^j y - n), m, n \in \mathbb{Z}\}$$

for  $j \in \mathbb{Z}$ .  $\bigcup_{j \in \mathbb{Z}} V_j$  is dense in  $\mathbb{R}^2$  and  $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$ .

Let  $\widehat{\psi}_k(\xi, \eta) = m_k(e^{i\xi/2}, e^{i\eta/2})\widehat{\phi}(\xi/2, \eta/2)$  and

$$W_{0,k} := \text{span}\{2^j \psi(2^j x - m, 2^j y - n), m, n \in \mathbb{Z}\}$$

for  $k = 1, 2, 3$ . In order for  $\psi_k$  to be orthonormal wavelets, we need to have

- 1)  $\psi_k$  is orthonormal for  $k = 1, 2, 3$  and
- 2)

$$V_1 = V_0 \oplus W_{0,1} \oplus W_{0,2} \oplus W_{0,3}.$$

These conditions 1) and 2) are equivalent to (58) (cf. [40]). Therefore, we mainly review recent methods for solving (57) and (58).

### Method of Tensor Product

Starting with univariate wavelets, we can use the method of tensor product to construct separable wavelets. Let us recall the method below.

Let  $P_a(x)$  and  $P_b(x)$  be two univariate Laurent polynomials satisfying  $P_a(1) = 1, P_b(1) = 1$ ,

$$|P_a(x)|^2 + |P_a(-x)|^2 = 1 \quad \text{and} \quad |P_b(x)|^2 + |P_b(-x)|^2 = 1.$$

Letting  $\widetilde{Q}_a = xP_a(-1/x)$  and  $\widetilde{Q}_b(x) = xP_b(-1/x)$  be the Laurent polynomials associated with  $P_a$  and  $P_b$ , respectively, we know

$$P_a(x)\overline{\widetilde{Q}_a(x)} + P_a(-x)\overline{\widetilde{Q}_a(-x)} = 0$$

$$P_b(y)\overline{\widetilde{Q}_b(y)} + P_b(-y)\overline{\widetilde{Q}_b(-y)} = 0.$$

With these relations, we define

$$m_0(x, y) = P_a(x)P_b(y),$$

$$m_1(x, y) = \widetilde{Q}_a(x)P_b(y),$$

$$m_2(x, y) = P_a(x)\widetilde{Q}_b(y),$$

$$m_3(x, y) = \widetilde{Q}_a(x)\widetilde{Q}_b(y).$$

Then,  $m_0$  satisfies (57) and  $m_j, j = 0, 1, 2, 3$  satisfy (58). This is an easy method. However, the method emphasizes horizontal and vertical directions which may not be desirable for applications. Therefore, these prompt for construction of nonseparable wavelets.

### The Ayache Method

In [1], Ayache proposed two methods for constructing bivariate nonseparable compactly supported orthonormal wavelets. The wavelets constructed using one of the methods is called semi-separable wavelets.

Starting with a separable filter  $m(x, y) = P_a(x)P_b(y)$ , we write

$$m(x, y) = P_a(x) \frac{P_b(y) + P_b(-y)}{2} + P_a(x) \frac{P_b(y) - P_b(-y)}{2}.$$

Let

$$(P_b)_e(y) = \frac{1}{2}(P_b(y) + P_b(-y)),$$

$$(P_b)_o(y) = \frac{1}{2}(P_b(y) - P_b(-y)).$$

be the even and odd part of  $P_b(x)$ . That is,  $m(x, y) = P_a(x)(P_b)_e(y) + P_a(x)(P_b)_o(y)$ . Let  $P_c(x)$  be another univariate CQF which is different from  $P_a(x)$ . We define

$$m_0(x, y) = P_a(x)(P_b)_e(y) + P_c(x)(P_b)_o(y)$$

$$m_1(x, y) = \widetilde{P}_a(x)(P_b)_e(y) + \widetilde{P}_c(x)(P_b)_o(y)$$

$$m_2(x, y) = P_a(x)(\widetilde{P}_b)_e(y) + P_c(x)(\widetilde{P}_b)_o(y)$$

$$m_3(x, y) = \widetilde{P}_a(x)(\widetilde{P}_b)_e(y) + \widetilde{P}_c(x)(\widetilde{P}_b)_o(y).$$

Then we have

**Lemma 16** *Let  $m_0(x, y)$  be a Laurent polynomial defined above. Then  $m_0(x, y)$  satisfies (57) and  $m_j, j = 0, 1, 2, 3$  satisfy (58).*

Let us give some simple examples.

**Example 23** Let  $P_a(x) = P_b(x) = \frac{1+x}{2}$  and  $P_c(x) = \frac{1+x^3}{2}$ . Then

$$m(x, y) = \frac{1+x}{4} + \frac{1+x^3}{4}y$$

is a simple bivariate nonseparable filter satisfying (57). It can be easily verified that  $m(x, y)$  satisfies the bivariate Cohen condition for the orthonormality. Thus,  $m(x, y)$  generates a bona fide scaling function  $\phi$ .

**Example 24** Let  $P_a(x) = P_b(x) = \frac{1+x}{2}$  and

$$P_c(x) = \frac{1 + \sqrt{3}}{8} + \frac{3 + \sqrt{3}}{8}x + \frac{3 - \sqrt{3}}{8}x^2 + \frac{1 - \sqrt{3}}{8}x^3.$$

Then  $m(x, y) = \frac{1+x}{4} + \frac{y}{2}P_c(x)$  is another simple bivariate nonseparable filter satisfying (57).

When  $P_a$  and  $P_c$  are masks associated with scaling functions  $\phi_a$  and  $\phi_c$ , if  $P_b$  is sufficiently close to  $P_a$ , then  $m_0(x, y)$  defined above will generate an orthonormal refinable function  $\phi$ . Indeed,  $m_0$  will satisfy the bivariate generalization of Cohen's orthonormal condition when  $P_b$  is sufficiently close to  $P_a$  (cf. [1] and [2]).

### The Maass–Ayache Method

This is the second method that Ayache proposed in [1]. Let  $\lambda(x, y)$  and  $\mu(x, y)$  be two even Laurent polynomials satisfying

$$|\lambda(x, y)|^2 + |\mu(x, y)|^2 = 1,$$

where  $\lambda(x, y)$  also satisfies  $\lambda(1, 1) = 1$ . We define

$$\begin{aligned} m_0(x, y) &= \lambda(x, y)P_a(x)P_b(y) + \mu(x, y)\widetilde{P}_a(x)P_b(y) \\ m_1(x, y) &= \overline{\mu(x, y)}P_a(x)P_b(y) - \overline{\lambda(x, y)}\widetilde{P}_a(x)P_b(y) \\ m_2(x, y) &= \lambda(x, y)P_a(x)\widetilde{P}_b(y) + \mu(x, y)\widetilde{P}_a(x)\widetilde{P}_b(y) \\ m_3(x, y) &= \overline{\mu(x, y)}P_a(x)\widetilde{P}_b(y) - \overline{\lambda(x, y)}\widetilde{P}_a(x)\widetilde{P}_b(y). \end{aligned}$$

**Lemma 17**  $m(x, y)$ ,  $m_1(x, y)$ ,  $m_2(x, y)$ ,  $m_3(x, y)$  so defined above satisfy the perfect reconstruction condition (58).

Next let  $q(x)$  be a nonzero Laurent polynomial with  $0 \leq q(x) \leq 1$  and  $q(1) = 0$ . We define

$$\lambda(x) = 1 - \frac{1}{4}q(x)$$

and  $v(x)$  such that

$$|\lambda(x)|^2 + |v(x)|^2 = 1. \quad (60)$$

We now define  $m_{N,0}(x, y)$  by

$$m_{N,0}(x, y) = D_N(y)(\lambda(x^2)\widehat{D}_N(x) + v(x^2)\widetilde{D}_N(x)), \quad (61)$$

where  $D_N$  is the filter associated with Daubechies' wavelet and  $\widetilde{D}_N(x) = xD_N(-x)$  is a conjugate filter of  $D_N(x)$ .

We refer [1] for the discussion of the orthonormality and regularity of the refinable function  $\phi_N$  generated by  $m_{N,0}$ .

Also, in [69], Maass proposed a similar method for  $m_0(x, y)$  such that the refinable function  $\phi$  associated with

$m_0$  satisfies the orthonormal condition under the dilation matrix  $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ : i. e.,

$$|m_0(x, y)|^2 + |m_0(-x, -y)|^2 = 1. \quad (62)$$

His method can be given as follows: Let  $\lambda$  and  $v$  be two even Laurent polynomials satisfying (60) with  $\lambda(1, 1) = 1$ . Then

$$m_0(x, y) = \lambda(x, y)D_N(x) + v(x, y)\widetilde{D}_N(x)$$

satisfies (62). Comparing with (61), we can see that the Maass method and the method Ayache used here are very similar. Thus, we call the construction the Maass–Ayache method.

### The Belogay and Wang Method

In [5], Belogay and Wang constructed nonseparable compactly supported orthonormal wavelets using dilation matrix  $\begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$ . However, their method can be modified to construct nonseparable wavelets using the dilation matrix  $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ . We begin with

**Lemma 18** Let  $P_a(x)$  and  $P_b(x)$  be two Laurent polynomials of  $x = e^{i\omega}$ . Suppose that  $P_a$  and  $P_b$  satisfy that  $P_a(1) = 1$ ,

$$|P_a(x)|^2 + |P_a(-x)|^2 + |P_b(x)|^2 + |P_b(-x)|^2 = 1 \quad (63)$$

and

$$m_a(x)m_b(1/x) + m_a(-x)m_b(-1/x) = 0. \quad (64)$$

Then, letting

$$m_0(x, y) = (P_a(x) + yP_b(x))(P_a(y) + x^2P_b(y)),$$

$m_0(1, 1) = 1$  and  $m_0(x, y)$  satisfies (57).

Next we define

$$\begin{aligned} m_1(x, y) &= y(P_a(x) + yP_b(x))(P_a(-1/y) \\ &\quad + x^2P_b(-1/y)), \\ m_2(x, y) &= x(P_a(-1/x) + 1/yP_b(-1/x)) \\ &\quad (P_a(y) + x^2P_b(y)), \\ m_3(x, y) &= xy(P_a(-1/x) + 1/yP_b(-1/x)) \\ &\quad (P_a(-1/y) + x^2P_b(-1/y)). \end{aligned}$$

Then it is easy to verify the following

**Lemma 19** Let  $m(x, y)$  and  $m_i(x, y)$ ,  $i = 1, 2, 3$  be Laurent polynomials defined above. They satisfy the perfect reconstruction condition (58).

Next we define these  $P_a$  and  $P_b$  which satisfy (63) and (64). Let  $T(x) = 2^{N-1}$  and  $S(x)$  such that

$$|S(x)|^2 = \sum_{k=0}^{N-1} \binom{N}{k} \left(\sin^2 \frac{\omega}{2}\right)^k \left(\cos^2 \frac{\omega}{2}\right)^{N-k} + \frac{3}{4} \left(\sin^2 \frac{\omega}{2}\right)^N$$

with  $x = e^{i\omega}$ . Then  $S$  and  $T$  satisfy

$$|S(x)|^2 + |T(1/x)|^2 \left| \frac{1-x}{4} \right|^{2N} = 1.$$

**Lemma 20** Let

$$P_a(x) = \left(\frac{1+x}{2}\right)^N \mathcal{L}_N(x) S(x^2) x^v$$

$$P_b(x) = \left(\frac{1+x}{2}\right)^N \left(\frac{1-x}{2}\right)^{2N} \mathcal{L}_N(-1/x) T(1/x^2),$$

where  $|\mathcal{L}_N(x)|^2 = \sum_{k=0}^{N-1} \binom{n-1+k}{k} \left(\sin^2 \frac{\omega}{2}\right)^k$ ,  $v$  is an integer such that  $v + N$  is odd. Then  $P_a(x)$  and  $P_b(x)$  satisfy (63) and (64).

This is the main step for constructing the Belogay–Wang nonseparable wavelets associated with dilation matrix  $2I_2$ . We omit the discussion on the orthonormality and regularity here. See their paper for details.

### The Karoui Method

In [43], Karoui proposed the following method to design nonseparable wavelets. Starting with a Laurent polynomial  $m(x, y)$  which satisfies

$$|m(x, y)|^2 + |m(-x, -y)|^2 = 1, \quad (65)$$

we define

$$m_0(x, y) = m(x, y) \overline{m(x/y, xy)}.$$

Then we have the following

**Lemma 21** Let  $m_0$  be defined above. Then  $m_0(x, y)$  satisfies (57) for  $|x| = 1$  and  $|y| = 1$ .

Next we define

$$m_1(x, y) = x/y m(x, y) \overline{m(-x/y, -xy)}$$

$$m_2(x, y) = x \overline{m(-x, -y)} m(x/y, xy),$$

$$m_3(x, y) = 1/y \overline{m(-x, -y)} m(-x/y, -xy).$$

Then we have

**Lemma 22** The four filters  $m_j$ ,  $j = 0, 1, 2, 3$  satisfy the perfect reconstruction condition (58).

Note that (65) is the same as (62) before. By the Maass method, we can construct  $m$  satisfy (65).

### The He and Lai Method

In [35], He and Lai construct many examples of nonseparable orthonormal wavelets. The constructive method starts with Laurent polynomial

$$m_0(x, y) = \sum_{0 \leq j \leq 3, 0 \leq k \leq 3} c_{j,k} x^j y^k$$

with  $x = e^{i\xi}$  and  $y = e^{i\eta}$ . First we write  $m(x, y)$  in its polyphase form:

$$m(x, y) = f_0(x^2, y^2) + x f_1(x^2, y^2) + y f_2(x^2, y^2) + xy f_3(x^2, y^2),$$

where

$$f_v(x, y) = a_v + b_v x + c_v y + d_v xy, \quad v = 0, 1, 2, 3.$$

**Lemma 23**  $m(x, y)$  satisfies (57) if and only if the following 5 nonlinear equations hold

$$\sum_{v=0}^3 (a_v b_v + c_v d_v) = 0, \quad \sum_{v=0}^3 (a_v c_v + b_v d_v) = 0,$$

$$\sum_{v=0}^3 a_v d_v = 0, \quad \sum_{v=0}^3 b_v c_v = 0,$$

and

$$\sum_{v=0}^3 (a_v^2 + b_v^2 + c_v^2 + d_v^2) = \frac{1}{4}.$$

The requirements  $m(1, 1) = 1$  and  $m(-1, y) = 0 = m(x, -1)$  imply the following 5 linear equations

$$\sum_{v=0}^3 a_v + b_v + c_v + d_v = 1,$$

$$a_v + b_v + c_v + d_v = \frac{1}{4}, \quad v = 0, 1, 2, 3.$$

We were able to find a complete solution for  $c_{j,k}$ ,  $0 \leq j, k \leq 3$ .

**Theorem 15** Let

$$m(x, y) = \sum_{0 \leq j \leq 3, 0 \leq k \leq 3} c_{j,k} x^j y^k = \frac{(1+x)(1+y)}{16} \\ \times (a_{00} + a_{10}x + a_{01}y + a_{11}xy + a_{20}x^2 \\ + a_{21}x^2y + a_{12}xy^2 + a_{22}x^2y^2 + a_{02}y^2) \quad (66)$$

with

$$\begin{cases} a_{00} = 1 + \sqrt{2}(\cos \alpha + \cos \beta) + 2 \cos \theta \cos \xi \\ a_{10} = \sqrt{2}(\sin \alpha - \cos \alpha) - 2 \cos \theta \cos \xi \\ \quad + 2 \cos \theta \sin \xi \\ a_{01} = \sqrt{2}(\sin \beta - \cos \beta) - 2 \cos \theta \cos \xi \\ \quad + 2 \sin \theta \cos \eta \\ a_{11} = 2(\cos \theta \cos \xi + \sin \theta \sin \eta - \cos \theta \sin \xi \\ \quad - \sin \theta \cos \eta) \\ a_{20} = 1 + \sqrt{2}(\cos \beta - \sin \alpha) - 2 \cos \theta \sin \xi \\ a_{02} = 1 + \sqrt{2}(\cos \alpha - \sin \beta) - 2 \sin \theta \cos \eta \\ a_{21} = \sqrt{2}(\sin \beta - \cos \beta) - 2 \sin \theta \sin \eta \\ \quad + 2 \cos \theta \sin \xi \\ a_{12} = \sqrt{2}(\sin \alpha - \cos \alpha) - 2 \sin \theta \sin \eta \\ \quad + 2 \sin \theta \cos \eta \\ a_{22} = 1 - \sqrt{2}(\sin \alpha + \sin \beta) + 2 \sin \theta \sin \eta. \end{cases}$$

Suppose that  $\alpha, \beta, \theta, \xi, \eta$  satisfy the following

$$\cos \theta \cos \xi + \cos \theta \sin \xi + \sin \theta \cos \eta + \sin \theta \sin \eta \\ = 2 \sin \left( \alpha + \frac{\pi}{4} \right) \sin \left( \beta + \frac{\pi}{4} \right). \quad (67)$$

Then  $m_0(x, y)$  satisfies (57). On the other hand, it  $m_0(x, y)$  satisfies (57), then the coefficients  $c_{ij}$  can be given in the above form and satisfy (67).

**Theorem 16** For any given  $m(x, y)$  in (66) satisfying (57), one can construct Laurant polynomials  $m_1, m_2$ , and  $m_3$  such that  $m_j, j = 0, 1, 2, 3$  satisfy the perfect reconstruction condition (58).

Their proof is constructive and has been implemented in MATLAB. See [35] for details. Let us give two examples of  $m(x, y)$ .

**Example 25** The following are two examples which has rational coefficients. One can construct  $m_j, j = 1, 2, 3$  associated with any of the following such that they satisfy the perfect reconstruction condition (58).

$$m(x, y) = \frac{(1+x)(1+y)}{100} (11 + 6x - 2x^2 + 6y \\ + 13xy - 4x^2y - 2y^2 - 4xy^2 + x^2y^2) \\ m(x, y) = \frac{(1+x)(1+y)}{3468} (544 + 120x - 52x^2 \\ + 120y + 416xy - 128x^2y - 52y^2 - 128xy^2 \\ + 27x^2y^2).$$

There are many other examples including those which are continuous. We refer to [35] for detail.

### Method of Symmetry

Another easy method to construct bivariate compactly supported wavelets is to use the method of symmetry. Suppose that a Laurent polynomial  $m_0(x, y)$  satisfies a symmetry property:  $m_0(1/x, 1/y) = x^{-M}y^{-N}m_0(x, y)$ . Suppose that  $m_0(x, y)$  satisfies (57). Then the other three filters  $m_j, j = 1, 2, 3$  can be easily obtained by

$$\begin{aligned} m_1(x, y) &= m_0(-x, y) \\ m_2(x, y) &= x \cdot m_0(x, -y) \\ m_3(x, y) &= x \cdot m_0(-x, -y). \end{aligned}$$

In [57], we found the complete solutions for

$$m_0(x, y) = \sum_{0 \leq j, k \leq 5} c_{j,k} x^j y^k$$

which satisfies the orthonormal condition (57), symmetry condition

$$m_0(1/x, 1/y) = x^{-5}y^{-5}m_0(x, y)$$

and low-pass feature  $m_0(-1, y) = m_0(x, -1) = 0$ . We leave the detail to [57].

### The Multiwavelet Method

Assume that  $\phi$  is a given compactly supported scaling function which generates an MRA. We now discuss how to construct compactly supported orthonormal wavelets  $\psi_j, j = 1, \dots, n$  with  $n \geq 3$  associated with  $\phi$ . For simplicity, let

$$m(\xi, \eta) = \sum_{0 \leq j \leq 5, 0 \leq k \leq 5} c_{j,k} e^{i(j\xi + k\eta)}.$$

Consider a function vector

$$\Phi(x, y) = \begin{bmatrix} 2\phi(2x, 2y) \\ 2\phi(2x - 1, 2y) \\ 2\phi(2x, 2y - 1) \\ 2\phi(2x - 1, 2y - 1) \end{bmatrix}.$$

Sine  $\phi$  is orthonormal, so is  $\Phi(x, y)$ , i. e.,

$$\int_{\mathbb{R}^2} \Phi(x - \ell, y - k) \Phi(x, y)^T dx dy = I_{4 \times 4} \delta_{0,\ell} \delta_{0,k}.$$

Writing  $\Phi = (\phi_1, \phi_2, \phi_3, \phi_4)^T$ , we let

$$\begin{aligned} \widetilde{V}_k &= \text{span}_{L_2} \{ \phi_j(2^k x - \ell, 2^k y - m), \\ &\quad \ell, m \in \mathbb{Z}, j = 1, 2, 3, 4 \}. \end{aligned}$$



Note that is  $\tilde{V}_0 = V_1$ . It follows that  $\{\tilde{V}_k, k \in Z\}$  forms an MRA and hence  $\Phi$  generates an MRA. It is clear that  $\Phi(x, y)$  is a refinable vector. That is,

$$\Phi(x, y) = \sum_{\ell, m \in Z} M_{\ell, m} \Phi(2x - \ell, 2y - m).$$

In terms of Fourier transform, we have

$$\hat{\Phi}(\xi, \eta) = M(\xi/2, \eta/2) \hat{\Phi}(\xi/2, \eta/2),$$

where  $M(\xi, \eta) = \frac{1}{4} \sum_{\ell, m} M_{\ell, m} e^{i(\ell\xi + m\eta)}$ .

It follows that

$$\sum_{\ell, m \in Z} \hat{\Phi}(\xi + 2\ell\pi, \eta + 2m\pi) \hat{\Phi}(\xi + 2\ell\pi, \eta + 2m\pi)^* = I_{4 \times 4}.$$

Thus, we have

$$\begin{aligned} & M(\xi, \eta) M(\xi, \eta)^* + M(\xi + \pi, \eta) M(\xi + \pi, \eta)^* \\ & + M(\xi, \eta + \pi) M(\xi, \eta + \pi)^* + M(\xi + \pi, \eta + \pi) \\ & M(\xi + \pi, \eta + \pi)^* = I_{4 \times 4}. \end{aligned}$$

Let  $\tilde{W}_k$  be the orthogonal complement of  $\tilde{V}_k$  in  $\tilde{V}_{k+1}$ . We will construct three compactly supported orthonormal multi-wavelet vectors  $\Psi_1, \Psi_2, \Psi_3 \in \tilde{W}_0$  such that

$$\tilde{V}_1 = \tilde{V}_0 \oplus \tilde{W}_0,$$

and for  $j, k = 1, 2, 3$ ,

$$\int_{\mathbb{R}^2} \Psi_j(x - \ell, y - m) \Psi_k(x, y)^T dx dy = 0, \quad j \neq k,$$

where

$$\tilde{W}_0 = \text{span}_{L_2} \{ \Psi_{j,k}(x - \ell, y - m), \ell, m \in Z, j = 1, 2, 3, k = 1, \dots, 4 \}$$

and  $\Psi_j = (\Psi_{j1}, \Psi_{j2}, \Psi_{j3}, \Psi_{j4})^T, j = 1, 2, 3$ . Writing

$$\hat{\Psi}_j(\xi, \eta) = M_j(\xi/2, \eta/2) \hat{\Phi}(\xi/2, \eta/2), \quad j = 1, 2, 3,$$

we need to find matrices  $M_j(\xi, \eta)$  with polynomial entries in  $(e^{i\xi}, e^{i\eta})$  such that

$$\begin{bmatrix} M(\xi, \eta) & M(\xi + \pi, \eta) \\ M_1(\xi, \eta) & M_1(\xi + \pi, \eta) \\ M_2(\xi, \eta) & M_2(\xi + \pi, \eta) \\ M_3(\xi, \eta) & M_3(\xi + \pi, \eta) \end{bmatrix}$$

$$\begin{bmatrix} M(\xi, \eta + \pi) & M(\xi + \pi, \eta + \pi) \\ M_1(\xi, \eta + \pi) & M_1(\xi + \pi, \eta + \pi) \\ M_2(\xi, \eta + \pi) & M_2(\xi + \pi, \eta + \pi) \\ M_3(\xi, \eta + \pi) & M_3(\xi + \pi, \eta + \pi) \end{bmatrix} \quad (68)$$

is a unitary matrix.

By the above properties and the fact that  $\Phi$  generates an MRA, we know that  $\{\Psi_{jk}, j = 1, 2, 3, k = 1, 2, 3, 4\}$  are compactly supported orthonormal wavelet functions for  $L_2(\mathbb{R}^2)$  if we have (68).

To solve this matrix extension problem, i.e., finding  $M_j, j = 1, 2, 3$  such that (68) holds, we need the following lemma whose proof is based on the ideas in [35].

**Theorem 17** Let  $M(\xi, \eta)$  be a matrix of size  $4 \times 4$  with polynomial entries in  $e^{i\xi}$  and  $e^{i\eta}$  with coordinate degrees  $\leq (3, 3)$ . Suppose that

$$\begin{aligned} I_{4 \times 4} &= M(\xi, \eta) M(\xi, \eta)^* + M(\xi + \pi, \eta) M(\xi + \pi, \eta)^* \\ &+ M(\xi, \eta + \pi) M(\xi, \eta + \pi)^* \\ &+ M(\xi + \pi, \eta + \pi) M(\xi + \pi, \eta + \pi)^*. \end{aligned}$$

Then there exist polynomial matrix  $M_j, j = 1, 2, 3$  such that the matrix in (68) is unitary.

Therefore we can have

**Theorem 18** Suppose that  $\phi(x, y) \in L_2(\mathbb{R}^2)$  is a scaling function associated with dilation matrix  $2I_2$ . Let

$$m(\xi, \eta) = \frac{1}{4} \sum_{\substack{0 \leq j \leq 5 \\ 0 \leq k \leq 5}} c_{jk} e^{i(j\xi + k\eta)}$$

be the mask associated with  $\phi$ . Then there exist 12 compactly supported orthonormal wavelets  $\psi_{jk}, j = 1, 2, 3, k = 1, 2, 3, 4$  such that translates dilation of these  $\psi_{j,k}$ 's form an orthonormal basis for  $L_2(\mathbb{R}^2)$ .

The detail of the proof is contained in [49]. What happens when the support size of  $\phi$  is bigger. Suppose that

$$m(\xi, \eta) = \sum_{\substack{0 \leq j \leq 9 \\ 0 \leq k \leq 9}} c_{jk} e^{i(j\xi + k\eta)}$$

is a trigonometric polynomial associated with a scaling function  $\phi$ . We will let

$$\Phi(x, y) = \begin{bmatrix} 4\phi(4x, 4y) \\ 4\phi(4x + 1, 4y) \\ 4\phi(4x, 4y + 1) \\ 4\phi(4x + 1, 4y + 1) \\ 4\phi(4x + 2, 4y) \\ 4\phi(4x + 2, 4y + 1) \\ 4\phi(4x + 1, 4y + 2) \\ 4\phi(4x, 4y + 2) \\ 4\phi(4x + 3, 4y) \\ 4\phi(4x + 3, 4y + 1) \\ 4\phi(4x + 3, 4y + 2) \\ 4\phi(4x + 3, 4y + 3) \\ 4\phi(4x + 2, 4y + 3) \\ 4\phi(4x + 1, 4y + 3) \\ 4\phi(4x, 4y + 3) \end{bmatrix}_{16 \times 1}.$$

Then  $\Phi = (\phi_1, \dots, \phi_{16})^T$  is an orthonormal scaling vector. Let

$$\widehat{V}_0 = \text{span}_{L_2} \{\phi_j(x - \ell, y - m), \ell, m \in \mathbb{Z}, j = 1, \dots, 16\}.$$

It is easy to see that  $\widehat{V}_0 = V_2$ . Thus,  $\Phi$  generates a bona fide MRA. The above construction procedure can be simply extended for this  $\Phi$ . We can construct three multi-wavelet vectors of size  $16 \times 1$ . The details can be found in [49]. Therefore, we conclude

**Theorem 19** Suppose that  $\phi(x, y) \in L_2(\mathbb{R}^2)$  is a scaling function associated with dilation matrix  $2I_2$ . Then there exist compactly supported orthonormal wavelets  $\psi_{jk}$ ,  $j = 1, 2, 3, k = 1, \dots, n$  with appropriate  $n$  dependent on the size of the support of  $\phi$  such that translates dilation of these  $\psi_{j,k}$ 's form an orthonormal basis for  $L_2(\mathbb{R}^2)$ .

### Biorthogonal Box Spline Wavelets

In a previous section, we have used B-spline function  $N_n$  to construct biorthogonal dual function  $\tilde{B}_n$  and the associated compactly supported biorthogonal wavelets. Since bivariate box splines are a natural generalization of B-spline functions, we shall present a constructive method to find biorthogonal wavelets associated with box spline functions.

Let  $B_{l,m,n}$  be the bivariate box spline function whose Fourier transform is

$$\widehat{B}_{l,m,n}(\omega_1, \omega_2) = \left( \frac{1 - e^{i\omega_1}}{i\omega_1} \right)^l \left( \frac{1 - e^{i\omega_2}}{i\omega_2} \right)^m \left( \frac{1 - e^{i(\omega_1 + \omega_2)}}{i(\omega_1 + \omega_2)} \right)^n.$$

(For properties of box spline functions, see [7,11,60] For computation of these bivariate box spline functions, see [47]). It is known that  $B_{l,m,n}$  generates a multi-resolution approximation of  $L_2(\mathbb{R}^2)$  (cf. [74]). We are interested in constructing a compactly supported function  $\tilde{B}_{l,m,n}$  generating a multi-resolution approximation of  $L_2(\mathbb{R}^2)$  which is a biorthogonal dual to  $B_{l,m,n}$  in the following sense:

$$\iint_{\mathbb{R}^2} B_{l,m,n}(x - j, y - k) \tilde{B}_{l,m,n}(x - j', y - k') dx dy = \delta_{j,j'} \delta_{k,k'} \quad (69)$$

for all integers  $j, k \in \mathbb{Z}$ , where  $\delta_{j,k}$  is the standard Kronecker notation defined by  $\delta_{j,k} = 0$  if  $j \neq k$  and  $\delta_{j,k} = 1$  if  $j = k$  and  $\mathbb{Z}$  is the collection of all integers.

We are furthermore interested in constructing compactly supported biorthogonal wavelets  $\psi_j$ ,  $j = 1, 2, 3$  and  $\tilde{\psi}_j$ ,  $j = 1, 2, 3$  and two families of FIR filters  $\{M_j, j = 0, 1, 2, 3\}$  and  $\{J_j, j = 0, 1, 2, 3\}$  with

$$\widehat{\psi}_j(\omega_1, \omega_2) = M_j(e^{i\frac{\omega_1}{2}}, e^{i\frac{\omega_2}{2}}) \widehat{B}_{l,m,n}\left(\frac{\omega_1}{2}, \frac{\omega_2}{2}\right), \quad j = 1, 2, 3, \quad (70)$$

and

$$\widehat{\tilde{\psi}}_j(\omega_1, \omega_2) = J_j\left(e^{i\frac{\omega_1}{2}}, e^{i\frac{\omega_2}{2}}\right) \widehat{\tilde{B}}_{l,m,n}\left(\frac{\omega_1}{2}, \frac{\omega_2}{2}\right), \quad j = 1, 2, 3, \quad (71)$$

such that the dilations and translates of the  $\psi_j$ 's and  $\tilde{\psi}_j$ 's form two dual Riesz bases for  $L_2(\mathbb{R}^2)$  and the two families form an exact reconstruction of synthesis/analysis filter bank for image/data processing.

In the following we mainly follow the construction of biorthogonal box spline wavelets in [36]. Denote  $z_1 = e^{i\omega_1}$  and  $z_2 = e^{i\omega_2}$ . Let

$$M_0(z_1, z_2) = \left( \frac{1 + z_1}{2} \right)^l \left( \frac{1 + z_2}{2} \right)^m \left( \frac{1 + z_1 z_2}{2} \right)^n$$

be a mask associated with the box spline function  $B_{l,m,n}$ . We look for a mask  $J_0(z_1, z_2)$  in the form

$$\overline{J_0(z_1, z_2)} = \left( \frac{1 + z_1}{2} \right)^{\tilde{n}-l} \left( \frac{1 + z_2}{2} \right)^{\tilde{n}-m} \times \left( \frac{1 + z_1 z_2}{2} \right)^{\tilde{m}-n} H(z_1, z_2) D(z_1 z_2) \quad (72)$$

with  $\tilde{n} > l$ ,  $\tilde{n} > m$  and odd integer  $\tilde{m} > n$  such that

$$\begin{aligned} M_0(z_1, z_2) \overline{J_0(z_1, z_2)} + M_0(-z_1, z_2) \overline{J_0(-z_1, z_2)} \\ + M_0(z_1, -z_2) \overline{J_0(z_1, -z_2)} \\ + M_0(-z_1, -z_2) \overline{J_0(-z_1, -z_2)} = 1. \end{aligned} \quad (73)$$

Recall from (14) that there exists a polynomial  $P_N(y)$  of degree  $< N$  such that

$$(1-y)^N P_N(y) + y^N P_N(1-y) = 1, \quad (74)$$

In fact, we have

$$P_N(y) = \sum_{k=0}^{N-1} \binom{2N-1}{k} (1-y)^{N-1-k} y^k.$$

**Theorem 20** Let  $\tilde{n} > n$  and  $\tilde{m} = 2\hat{m} + 1$ . Let  $J_0(z_1, z_2)$  be defined in (72) with  $H$  and  $D$  defined by

$$\begin{aligned} H(z_1, z_2) = \sum_{k=0}^{\tilde{n}-1} \binom{2\tilde{n}-1}{k} \left( \frac{1+z_1}{2} \frac{1+z_2}{2} \right)^{\tilde{n}-1-k} \\ \left( \frac{1-z_1}{2} \frac{1-z_2}{2} \right)^k, \end{aligned} \quad (75)$$

and

$$D(e^{i(\omega_1+\omega_2)}) = e^{-i(\omega_1+\omega_2)N} P_{\tilde{n}+\tilde{m}} \left( \sin^2 \left( \frac{\omega_1+\omega_2}{2} \right) \right). \quad (76)$$

Then  $J_0$  is a dual of  $M_0$  satisfying (73).

*Proof* We first note that

$$\frac{1+z_1 z_2}{2} = \frac{1+z_1}{2} \frac{1+z_2}{2} + \frac{1-z_1}{2} \frac{1-z_2}{2}.$$

By letting  $H(z_1, z_2)$  be defined in (75), we have, similar to the derivation of  $P_N(y)$  in (14)

$$\begin{aligned} \left( \frac{1+z_1 z_2}{2} \right)^{2\tilde{n}-1} \\ = \left( \frac{1+z_1}{2} \frac{1+z_2}{2} + \frac{1-z_1}{2} \frac{1-z_2}{2} \right)^{2\tilde{n}-1} \\ = \sum_{k=0}^{\tilde{n}-1} \binom{2\tilde{n}-1}{k} \end{aligned}$$

$$\begin{aligned} \cdot \left( \frac{1+z_1}{2} \frac{1+z_2}{2} \right)^{2\tilde{n}-1-k} \left( \frac{1-z_1}{2} \frac{1-z_2}{2} \right)^k \\ + \sum_{\ell=0}^{\tilde{n}-1} \binom{2\tilde{n}-1}{2\tilde{n}-1-\ell} \left( \frac{1+z_1}{2} \frac{1+z_2}{2} \right)^\ell \\ \cdot \left( \frac{1-z_1}{2} \frac{1-z_2}{2} \right)^{2\tilde{n}-1-\ell} = \left( \frac{1+z_1}{2} \frac{1+z_2}{2} \right)^{\tilde{n}} \\ \cdot H(z_1, z_2) + \left( \frac{1-z_1}{2} \frac{1-z_2}{2} \right)^{\tilde{n}} H(-z_1, -z_2) \end{aligned}$$

and similarly,

$$\begin{aligned} \left( \frac{1-z_1 z_2}{2} \right)^{2\tilde{n}-1} = \left( \frac{1-z_1}{2} \frac{1+z_2}{2} \right. \\ \left. + \frac{1+z_1}{2} \frac{1-z_2}{2} \right)^{2\tilde{n}-1} = \left( \frac{1-z_1}{2} \frac{1+z_2}{2} \right)^{\tilde{n}} \\ \cdot H(-z_1, z_2) + \left( \frac{1+z_1}{2} \frac{1-z_2}{2} \right)^{\tilde{n}} H(z_1, -z_2). \end{aligned}$$

With the definition of  $J_0$  in (73), (73) may be simplified as follows:

$$\begin{aligned} M_0(z_1, z_2) \overline{J_0(z_1, z_2)} + M_0(-z_1, -z_2) \overline{J_0(-z_1, -z_2)} \\ + M_0(-z_1, z_2) \overline{J_0(-z_1, z_2)} + M_0(z_1, -z_2) \overline{J_0(z_1, -z_2)} \\ = \left[ \left( \frac{1+z_1}{2} \frac{1+z_2}{2} \right)^{\tilde{n}} H(z_1, z_2) + \left( \frac{1-z_1}{2} \frac{1-z_2}{2} \right)^{\tilde{n}} \right. \\ \left. H(-z_1, -z_2) \right] \times \left( \frac{1+z_1 z_2}{2} \right)^{\tilde{m}} D(z_1 z_2) \\ + \left[ \left( \frac{1-z_1}{2} \frac{1+z_2}{2} \right)^{\tilde{n}} H(-z_1, z_2) + \left( \frac{1+z_1}{2} \frac{1-z_2}{2} \right)^{\tilde{n}} \right. \\ \left. H(z_1, -z_2) \right] \times \left( \frac{1-z_1 z_2}{2} \right)^{\tilde{m}} D(-z_1 z_2) \\ = \left( \frac{1+z_1 z_2}{2} \right)^{2\tilde{n}+\tilde{m}-1} D(z_1 z_2) + \left( \frac{1-z_1 z_2}{2} \right)^{2\tilde{n}+\tilde{m}-1} \\ D(-z_1 z_2). \end{aligned}$$

Let  $\tilde{m} = 2\hat{m} + 1$  and  $N = \tilde{n} + \hat{m}$ . Recall  $z_1 = e^{i\omega_1}$  and  $z_2 = e^{i\omega_2}$ . Then the last equation may be simplified further:

$$\begin{aligned} \left( \cos^2 \frac{\omega_1 + \omega_2}{2} \right)^N e^{i(\omega_1+\omega_2)N} D(e^{i(\omega_1+\omega_2)}) \\ + \left( \sin^2 \frac{\omega_1 + \omega_2}{2} \right)^N (-1)^N e^{i(\omega_1+\omega_2)N} D(-e^{i(\omega_1+\omega_2)}). \end{aligned}$$

Let  $y = \sin^2 \left( \frac{\omega_1+\omega_2}{2} \right)$  and recognize that  $e^{i(\omega_1+\omega_2)N} D(e^{i(\omega_1+\omega_2)}) = P_N(y)$ . We can see that the above equa-

tion is just the left-hand side of (74). Therefore, we have established the results of Theorem 20.  $\square$

We remark here that the filter  $J_0(z_1, z_2)$  is a linear phase filter. It is known that the Fourier transform of box spline  $B_{l,m,n}$  is

$$\widehat{B}_{l,m,n}(\omega_1, \omega_2) = \prod_{k=1}^{\infty} M_0\left(e^{i\frac{\omega_1}{2^k}}, e^{i\frac{\omega_2}{2^k}}\right) \in L_2(\mathbb{R}^2).$$

We now construct the dual functions  $\tilde{B}_{l,m,n}$  in terms of its Fourier transform by

$$\widehat{\tilde{B}}_{l,m,n}(\omega_1, \omega_2) = \prod_{k=1}^{\infty} J_0\left(e^{i\frac{\omega_1}{2^k}}, e^{i\frac{\omega_2}{2^k}}\right).$$

To see  $\tilde{B}_{l,m,n}$  is in  $L_2(\mathbb{R}^2)$  and generates a MRA in the bi-variate setting, we have the following (cf. [36])

**Theorem 21** *Let  $\tilde{n}$  and  $\tilde{m}$  be large enough. Then  $\tilde{B}_{l,m,n}$  is a well-defined compactly supported  $L_2$  function. Furthermore, for any  $\alpha > 0$ ,  $\tilde{B}_{l,m,n} \in C^\alpha(\mathbb{R}^2)$  if  $\tilde{n}$  and  $\tilde{m}$  sufficiently large, e. g.,*

$$\begin{aligned} \tilde{n} &> \frac{2(\max(l, m) + 2 + \alpha)}{2 - \log(3)/\log(2)}, \\ \tilde{m} &> 2 \frac{\tilde{n} \log(3)/\log(2) + n - 1}{2 - \log(3)/\log(2)} + 1. \end{aligned}$$

We next show that  $\tilde{B}_{l,m,n}$  defined above is a biorthogonal dual to  $B_{l,m,n}$  in the sense of (69). We first see that  $\widehat{\tilde{B}}_{l,m,n}$  is continuous and  $\widehat{\tilde{B}}_{l,m,n}(0, 0)\widehat{B}_{l,m,n}(0, 0) = 1$ . It is straightforward to prove

**Theorem 22** *For any sufficiently large integers  $\tilde{n}$  and  $\tilde{m}$ ,*

$$\begin{aligned} \sum_{\ell \in \mathbb{Z}^2} \left| \widehat{\tilde{B}}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell) \widehat{B}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell) \right|^2 \\ \geq C_2 > 0. \end{aligned}$$

The same arguments in the proof of Theorem 22 can also show that

$$\sum_{\ell \in \mathbb{Z}^2} \left| \widehat{\tilde{B}}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell) \right|^2 \geq C_1. \quad (77)$$

It follows from Theorem 21 that

$$\sum_{\ell \in \mathbb{Z}^2} \left| \widehat{\tilde{B}}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell) \right|^2 \leq C_2. \quad (78)$$

Thus, letting

$$V_0 = \text{span}\{\tilde{B}_{l,m,n}(x - j, y - k), (j, k) \in \mathbb{Z}^2\},$$

the inequalities (77) and (78) imply that  $\{\tilde{B}_{l,m,n}(x - j, y - k), (j, k) \in \mathbb{Z}^2\}$  is a Riesz basis for  $V_0$ . Letting  $V_k := \{f(x/2^k, y/2^k) : \forall f(x, y) \in V_0\}$  for  $k \in \mathbb{Z}$ , we can show that  $\bigcup_k V_k$  is dense in  $L_2(\mathbb{R}^2)$  and  $\bigcap_k V_k = \{0\}$ . We leave the detail to the interested reader. Thus, we conclude that  $\tilde{B}_{l,m,n}$  generates a multi-resolution approximation of  $L_2(\mathbb{R}^2)$ . These complete the proof of the following

**Theorem 23** *Let  $\tilde{n}$  and  $\tilde{m}$  be sufficiently large. Then  $\tilde{B}_{l,m,n}$  generates a multi-resolution approximation of  $L_2(\mathbb{R}^2)$ . Also,  $\tilde{B}_{l,m,n}$  is a biorthogonal dual of  $B_{l,m,n}$ .*

Next we work on constructing biorthogonal wavelets associated with  $B_{l,m,n}$  and  $\tilde{B}_{l,m,n}$ . As before we start with the construction of two families of Laurent polynomials satisfying the following

$$\begin{aligned} \begin{bmatrix} M_0(z_1, z_2) & M_1(z_1, z_2) \\ M_0(-z_1, z_2) & M_1(-z_1, z_2) \\ M_0(z_1, -z_2) & M_1(z_1, -z_2) \\ M_0(-z_1, -z_2) & M_1(-z_1, -z_2) \end{bmatrix} \\ \begin{bmatrix} M_2(z_1, z_2) & M_3(z_1, z_2) \\ M_2(-z_1, z_2) & M_3(-z_1, z_2) \\ M_2(z_1, -z_2) & M_3(z_1, -z_2) \\ M_2(-z_1, -z_2) & M_3(-z_1, -z_2) \end{bmatrix} \\ \times \begin{bmatrix} J_0(z_1, z_2) \\ J_1(z_1, z_2) \\ J_2(z_1, z_2) \\ J_3(z_1, z_2) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (79) \end{aligned}$$

For convenience, let us denote by  $A(M_0, M_1, M_2, M_3)$  the coefficient matrix in (79). In order to have compactly supported wavelets we need  $M_j, J_j, j = 1, 2, 3$  to be Laurent polynomials in  $(z_1, z_2)$  and the matrix  $A(M_0, M_1, M_2, M_3)$  must have a nonzero monomial determinant, i. e.,  $Cz_1^l z_2^k$ .

To this end, we rewrite  $M_j, j = 0, 1, 2, 3$  in its poly-phase form

$$\begin{aligned} M_j(z_1, z_2) &= f_{j0}(z_1^2, z_2^2) + z_1 f_{j1}(z_1^2, z_2^2) \\ &\quad + z_2 f_{j2}(z_1^2, z_2^2) + z_1 z_2 f_{j3}(z_1^2, z_2^2). \end{aligned}$$

Similarly we have

$$\begin{aligned} M_j(-z_1, z_2) &= f_{j0}(z_1^2, z_2^2) - z_1 f_{j1}(z_1^2, z_2^2) \\ &\quad + z_2 f_{j2}(z_1^2, z_2^2) - z_1 z_2 f_{j3}(z_1^2, z_2^2) \\ M_j(z_1, -z_2) &= f_{j0}(z_1^2, z_2^2) + z_1 f_{j1}(z_1^2, z_2^2) \\ &\quad - z_2 f_{j2}(z_1^2, z_2^2) - z_1 z_2 f_{j3}(z_1^2, z_2^2) \\ M_j(-z_1, -z_2) &= f_{j0}(z_1^2, z_2^2) - z_1 f_{j1}(z_1^2, z_2^2) \\ &\quad - z_2 f_{j2}(z_1^2, z_2^2) + z_1 z_2 f_{j3}(z_1^2, z_2^2). \end{aligned}$$

We can easily check

$$\begin{aligned}
 & A(M_0, M_1, M_2, M_3) \\
 &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & z_1 & 0 & 0 \\ 0 & 0 & z_2 & 0 \\ 0 & 0 & 0 & z_1 z_2 \end{bmatrix} \\
 &\quad \times \begin{bmatrix} f_{00} & f_{10} & f_{20} & f_{30} \\ f_{01} & f_{11} & f_{21} & f_{31} \\ f_{02} & f_{12} & f_{22} & f_{32} \\ f_{03} & f_{13} & f_{23} & f_{33} \end{bmatrix}, \quad (80)
 \end{aligned}$$

where  $f_{jk} := f_{jk}(z_1^2, z_2^2)$ 's. It is easy to see that for a given  $M_0$ , the existence of the matrix  $A(M_0, M_1, M_2, M_3)$  such that its determinant is a monomial  $Cz_1^{2\mu}z_2^{2\nu}$  is equivalent to the existence of  $[f_{jk}]_{0 \leq j, k \leq 3}$  whose determinant is a monomial.

It is clear from the expression of  $M_0(z_1, z_2)$  associated with box spline  $B_{1,m,n}$  that  $M_0(z_1, z_2)$ ,  $M_0(-z_1, z_2)$ ,  $M_0(z_1, -z_2)$ ,  $M_0(-z_1, -z_2)$  have no common zeros in  $\mathbb{C}^2$ , where  $\mathbb{C}$  denotes the usual complex space. It follows that  $f_{00}, f_{01}, f_{02}, f_{03}$  have no common zeros.

We further show that the first three polyphase terms  $f_{00}(z_1^2, z_2^2)$ ,  $f_{01}(z_1^2, z_2^2)$ ,  $f_{02}(z_1^2, z_2^2)$  have no common zero in  $(\mathbb{C})^2$  (cf. [36]).

**Lemma 24** Suppose that  $f_{0j}$ ,  $j = 0, \dots, 3$  are polynomials in  $(z_1, z_2)$ . Suppose that  $f_{00}, f_{01}, f_{02}$  have no common zeros in  $(\mathbb{C})^2$ . Then there exist  $f_{k,j}$ ,  $j = 0, 1, 2, 3$  and  $k = 1, 2, 3$  such that the matrix  $[f_{k,j}]_{0 \leq k, j \leq 3}$  is of determinant  $\pm 1$ .

*Proof* By the well-known Hilbert Nullstellensatz there exist polynomials  $p_0, p_1, p_2$  such that  $p_0 f_{00} + p_1 f_{01} + p_2 f_{02} = 1$ . Then it is easy to check that

$$\begin{aligned}
 & \begin{bmatrix} f_{00} & 1 & 0 & 0 \\ f_{01} & 0 & 1 & 0 \\ f_{02} & 0 & 0 & 1 \\ f_{03} & -p_0(1-f_{03}) & -p_1(1-f_{03}) & -p_2(1-f_{03}) \end{bmatrix} \\
 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -p_0(1-f_{03}) & -p_1(1-f_{03}) & -p_2(1-f_{03}) & 1 \end{bmatrix} \\
 &\quad \times \begin{bmatrix} f_{00} & 1 & 0 & 0 \\ f_{01} & 0 & 1 & 0 \\ f_{02} & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}
 \end{aligned}$$

which is obviously of determinant  $-1$ . We choose this matrix for  $[f_{k,j}]_{0 \leq k, j \leq 3}$ .  $\square$

By Lemma above, for  $M_0$ , we can find  $\tilde{M}_1, \tilde{M}_2, \tilde{M}_3$  such that  $A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3)$  has a determinant which is

a nonzero monomial  $Cz_1^{2j'}z_2^{2k'}$ . A computation of the determinant of matrix  $A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3)$  from the right-hand side of (80) yields the determinant is of even power, i.e.,  $j = 2j'$  and  $k = 2k'$ . Without loss of generality, we may simply assume

$$\det(A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3)) = 1$$

by absorbing  $Cz_1^{2j'}z_2^{2k'}$  into  $\tilde{M}_1$ . Let us invert the matrix  $(A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3))^T$ . From the definition of the inverse matrix, we know there exist polynomial entries  $\bar{J}_0, \bar{J}_1, \bar{J}_2$ , and  $\bar{J}_3$  such that

$$(A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3))^T)^{-1} = A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$$

or equivalently,

$$(A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3))^{-1} = (A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3))^T.$$

Since the determinant is 1, we know that, by Cramer's rule,  $M_0$  is equal to the cofactor of  $\bar{J}_0$  in matrix  $A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$ . In particular, we have

$$\begin{aligned}
 M_0(z_1, z_2) = \det \begin{bmatrix} \bar{J}_1(-z_1, z_2) & \bar{J}_2(-z_1, z_2) \\ \bar{J}_1(z_1, -z_2) & \bar{J}_2(z_1, -z_2) \\ \bar{J}_1(-z_1, -z_2) & \bar{J}_2(-z_1, -z_2) \\ \bar{J}_3(-z_1, z_2) \\ \bar{J}_3(z_1, -z_2) \\ \bar{J}_3(-z_1, -z_2) \end{bmatrix} \quad (81)
 \end{aligned}$$

Note that expanding according to the first column of  $A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$  and by using the definition of the inverse matrix, we have

$$\begin{aligned}
 1 &= \det(A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)) \\
 &= \bar{J}_0(z_1, z_2)M_0(z_1, z_2) + \bar{J}_0(-z_1, z_2)M_0(-z_1, z_2) \\
 &\quad + \bar{J}_0(z_1, -z_2)M_0(z_1, -z_2) + \bar{J}_0(-z_1, -z_2) \\
 &\quad M_0(-z_1, -z_2).
 \end{aligned}$$

Replacing the first column of matrix  $A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$  by a column  $[\bar{J}_0(z_1, z_2), \bar{J}_0(-z_1, z_2), \bar{J}_0(z_1, -z_2), \bar{J}_0(-z_1, -z_2)]^T$  with  $\bar{J}_0$  being defined in (72), we get a new matrix  $A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$  whose determinant is

$$\begin{aligned}
 \det(A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)) &= \bar{J}_0(z_1, z_2)M_0(z_1, z_2) \\
 &\quad + \bar{J}_0(-z_1, z_2)M_0(-z_1, z_2) + \bar{J}_0(z_1, -z_2)M_0(z_1, -z_2) \\
 &\quad + \bar{J}_0(-z_1, -z_2)M_0(-z_1, -z_2) = 1
 \end{aligned}$$

by (73). We compute the inverse of  $A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$  and write

$$A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)^{-1} = A(q_0, M_1, M_2, M_3)^T.$$



By the definition of the inverse matrices, it is now easy to recognize that  $q_0 = M_0$  since (81). That is, we have

$$A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)A(M_0, M_1, M_2, M_3)^T = I_4$$

where  $I_4$  stands for the identity matrix of  $4 \times 4$ . Hence,

$$A(M_0, M_1, M_2, M_3)A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)^T = I$$

which implies (79). Therefore, we have obtained the following

**Theorem 24** Let  $M_0(z_1, z_2) = \left(\frac{1+z_1}{2}\right)^l \left(\frac{1+z_2}{2}\right)^m \left(\frac{1+z_1 z_2}{2}\right)^n$  and  $J_0$  defined in (72). Then there exist  $M_1, M_2, M_3$  and  $J_1, J_2, J_3$  such that the exact reconstruction condition (79) holds.

We are finally ready to define biorthogonal wavelet functions  $\psi_j$  and  $\tilde{\psi}_j$  in terms of Fourier transform by

$$\widehat{\psi}_j(\omega_1, \omega_2) = J_j \left( e^{i\frac{\omega_1}{2}}, e^{i\frac{\omega_2}{2}} \right) \widehat{B}_{l,m,n} \left( \frac{\omega_1}{2}, \frac{\omega_2}{2} \right),$$

and

$$\widehat{\tilde{\psi}}_j(\omega_1, \omega_2) = M_j \left( e^{i\frac{\omega_1}{2}}, e^{i\frac{\omega_2}{2}} \right) \widehat{B}_{l,m,n} \left( \frac{\omega_1}{2}, \frac{\omega_2}{2} \right),$$

for  $j = 1, 2, 3$ . By using a generalization of the proof in [22] we can prove

**Theorem 25** Let  $\psi_j$ ,  $j = 1, 2, 3$  and  $\tilde{\psi}_j$ ,  $j = 1, 2, 3$  be defined above. Let

$$\begin{aligned} \psi_{j,k,(\ell_1, \ell_2)}(x, y) &= 2^{-k} \psi_j(2^{-k}x - \ell_1, 2^{-k}y - \ell_2) \\ \tilde{\psi}_{j,k,(\ell_1, \ell_2)}(x, y) &= 2^{-k} \tilde{\psi}_j(2^{-k}x - \ell_1, 2^{-k}y - \ell_2) \end{aligned}$$

for  $(\ell_1, \ell_2) \in \mathbb{Z}^2$ ,  $k \in \mathbb{Z}$ , and  $j = 1, 2, 3$ . Then the  $\psi_{j,k,(\ell_1, \ell_2)}$ 's and  $\tilde{\psi}_{j,k,(\ell_1, \ell_2)}$ 's constitute two dual Riesz bases of  $L_2(\mathbb{R}^2)$ . They satisfy biorthogonal conditions.

$$\begin{aligned} \int_{\mathbb{R}^2} \psi_{j,k,(\ell_1, \ell_2)}(x, y) \tilde{\psi}_{i,m,(\ell_1, \ell_2)}(x, y) dx dy \\ = \delta_{i,j} \delta_{k,m} \delta_{\ell_1, n_1} \delta_{\ell_2, n_2}, \end{aligned}$$

for all  $i, j = 1, 2, 3$ ,  $k, m \in \mathbb{Z}$ , and  $\ell_1, \ell_2, n_1, n_2 \in \mathbb{Z}$ . That is,  $\psi_j$  and  $\tilde{\psi}_j$  are biorthogonal wavelets.

Although the above theory is beautiful, examples are difficult to give and hence, are omitted here. Trivariate biorthogonal box spline wavelets were discussed in [37]. The interested reader may refer to [37] for details.

## Multivariate Prewavelets

The general construction of prewavelets in the univariate setting (see Sect. "Prewavelets") can easily be generalized to the multivariate setting. We shall outline such a generalization here.

We start the definition of multi-resolution approximation of  $L_2(\mathbb{R}^d)$ .

**Definition 5** A multi-resolution approximation (MRA) of  $L_2(\mathbb{R}^d)$  is a sequence of subspaces  $V_j$ ,  $j \in \mathbb{Z}$  of  $L_2(\mathbb{R}^d)$  such that

- (i)  $V_j \subset V_{j+1}$ ;
- (ii)  $\bigcup_{j=-\infty}^{\infty} V_j$  is dense in  $H^s(\mathbb{R}^d)$ ;
- (iii)  $\bigcap_{j=-\infty}^{\infty} V_j = \{0\}$ ;
- (iv) there is a function  $\phi \in V_0$  such that the integer translates,  $\phi(x - m)$ ,  $m \in \mathbb{Z}^d$  form a Riesz basis for  $V_0$ , i. e., there exist two positive numbers  $A$  and  $B$  such that

$$\begin{aligned} A \left\| \{c_m, m \in \mathbb{Z}^d\} \right\|_2^2 \\ \leq \left\| \sum_{m \in \mathbb{Z}^d} c_m \phi_j(x - m) \right\|_2^2 \leq B \left\| \{c_m, m \in \mathbb{Z}^d\} \right\|_2^2, \end{aligned}$$

for all square summable sequence  $\{c_m, m \in \mathbb{Z}^d\}$ .

We shall say  $\phi$  generates a multiresolution approximation of  $L_2(\mathbb{R}^d)$  if letting  $V_j = \text{span}\{\phi(2^j x - k), k \in \mathbb{Z}^d\}$ ,  $V_j$  is a nested subspace satisfying the above conditions (i)–(iv). Next we need

**Definition 6** A collection  $\psi_{j,k}$ ,  $k = 2, \dots, 2^d$  of functions in  $L_2(\mathbb{R}^d)$  satisfying the following five properties are called prewavelets:

1. the closure  $W_k$  of the linear span of integer translates of  $\psi_k$  is orthogonal to the closure  $V_0$  of the linear span of integer translates of  $\phi$ ;
2.  $W_k$  is orthogonal each other among  $k = 2, \dots, 2^d$ ,
3.  $V_1$  is the direct sum  $V_0$  and  $W_k$ ,  $k = 2, \dots, 2^d$ ;
4. the integer translates of  $\psi_k$  form a Riesz basis for  $W_k$ ; That is, there exist two positive constants  $A$  and  $B$  such that

$$\begin{aligned} A \sum_{k=2}^{2^d} \sum_{m \in \mathbb{Z}^d} |c_{k,m}|^2 &\leq \left\| \sum_{k=2}^{2^d} \sum_{m \in \mathbb{Z}^d} c_{k,m} 2^{jd/2} \psi_k(2^j \cdot - m) \right\|_2^2 \\ &\leq B \sum_{k=2}^{2^d} \sum_{m \in \mathbb{Z}^d} |c_{k,m}|^2 \end{aligned}$$

for all square summable sequence  $\{c_{k,m}, k = 2, \dots, 2^d, m \in \mathbb{Z}^d\}$ .

In this section, we assume that there exists a compactly supported function  $\phi$  which generates an MRA of  $L_2(\mathbb{R}^d)$ . Let  $P$  be the mask polynomial defined by

$$\widehat{\phi}(2\omega) = P(z)\widehat{\phi}(\omega),$$

where  $\widehat{\phi}$  denotes the Fourier transform of  $\phi$  and  $z = \exp(i\omega)$ . We are looking for compactly supported functions  $\psi_k, k = 2, \dots, 2^d$  in  $V_1$  such that

$$V_1 = V_0 \bigoplus_{k=2}^{2^d} W_k,$$

and  $\phi(\cdot - m), \psi_k(\cdot - m), m \in \mathbb{Z}^d, k = 2, \dots, 2^d$  form a stable basis for  $V_1$ , where  $W_k$  is the closure of the linear span of integer translates of  $\psi_k(x - m), m \in \mathbb{Z}^d$ .

To do so, we first introduce a Laurent polynomial

$$\Phi(z) := \sum_{m \in \mathbb{Z}^d} \left\langle 2^{d/2} \phi(2x), 2^{d/2} \phi_j(2x - m) \right\rangle z^m.$$

This function  $\Phi$  is called the generalized Euler–Frobenius polynomial.

Next we need a necessary and sufficient condition for the orthogonality. Writing

$$g_k(2x) = \sum_{m \in \mathbb{Z}^d} c_{k,m} 2^{d/2} \phi(2x - m) \in V_1,$$

and

$$G_k(z) = \frac{1}{2^{d/2}} \sum_{m \in \mathbb{Z}^d} c_{k,m} z^m,$$

the Fourier transforms of  $g_k$  and  $\phi$  are related by

$$\widehat{g}_k(2\omega) = G_k(z)\widehat{\phi}(\omega).$$

Let

$$G_k = \text{closure}_{L_2(\mathbb{R}^d)} \{g_k(x - m), m \in \mathbb{Z}^d\}$$

be the closure of the linear span of integer translates of  $g_k$ .

Recall a special operator  $E$  which maps any Laurent polynomial  $f$  into such a Laurent polynomial  $E(f)$  which contains all the even index terms of  $f$ . For example, when  $d = 2$  and  $z = (z_1, z_2)$ ,

$$E(f(z)) = \frac{1}{4} (f(z_1, z_2) + f(-z_1, z_2) + f(z_1, -z_2) + f(-z_1, -z_2)).$$

One important property is

$$E(P(z)\overline{P(z)}\Phi(z)) = 2^{-d}\Phi(z^2)$$

(see a proof in [19].)

We have the following generalization of Theorem 8.

**Theorem 26**  $G_k$  is orthogonal to  $G_{k'}$  for  $k' \neq k$  if and only if

$$E(G_k(z)\overline{G_{k'}(z)}\Phi(z)) = 0. \quad (82)$$

We divide the construction of compactly supported pre-wavelets into two steps. The first step is to construct compactly supported  $g_k \in V_1, k = 1, \dots, 2^d$  such that the closure  $G_k$  of the linear span of integer translates  $g_k$  is orthogonal to  $V_0$  for each  $k$ . The second step is to use a technique like Gram–Schmidt orthonormal procedure to orthogonalize these  $g_k$ 's for different  $k$ .

To be more precise, we let  $\{n_1, \dots, n_{2^d}\} = \{0, 1\}^d$  with  $n_k \in \mathbb{Z}^{2^d}$  and  $g_k \in V_1$  satisfy

$$g_k(x - m) \perp V_0, \quad m \in \mathbb{Z}^d$$

and

$$\begin{aligned} & 2^{d/2} \phi(2x - n_k) \\ &= \sum_{m \in \mathbb{Z}^d} (a_{k,m} 2^{d/2} \phi(x - m) + b_{k,m} 2^{d/2} g_k(2^j x - m)) \end{aligned}$$

for each  $k \in \{1, \dots, 2^d\}$ . That is, we want to have  $G_k$  is orthogonal to  $V_0$  and  $V_1 = V_0 \bigoplus (G_1 + \dots + G_{2^d})$ . In terms of Fourier transform, the above equations can be rewritten as

$$\begin{aligned} \frac{1}{2^{d/2}} e^{in_k \omega/2} \widehat{\phi}\left(\frac{\omega}{2}\right) &= A_k(\omega) \widehat{\phi}(\omega) + B_k(\omega) \widehat{g}_k(\omega) \\ &= A_k(\omega) P(\omega/2) \widehat{\phi}(\omega/2) + B_k(\omega) G_k(\omega/2) \widehat{\phi}(\omega/2), \end{aligned}$$

where  $A_k(\omega) = \sum_{m \in \mathbb{Z}^d} 2^{d/2} a_{k,m} e^{im\omega}$  and  $B_k(\omega) = \sum_{m \in \mathbb{Z}^d} 2^{d/2} b_{k,m} e^{im\omega}$ . Here, we have abused the notation of  $P(z) G_k(z)$ , that is, we use  $P(\omega)$  instead of  $P(z)$  and  $G_k(\omega)$  instead of  $G_k(z)$  with  $z = e^{i\omega}$  just for convenience. It follows that

$$A_k(2\omega)P(z) + B_k(2\omega)G_k(z) = \frac{e^{in_k \omega}}{2^{d/2}}, \quad (83)$$

for  $k = 1, \dots, 2^d$ . Using Theorem 26, the solution of  $A_k, B_k$  and  $G_k$  can be easily found as shown in the following.

**Lemma 25** Suppose that  $\Phi(z^2) = 2^d E(P(z)\overline{P(z)}\Phi(z)) \neq 0$  for all  $z$  in torus  $T^d$ . Let

$$\begin{aligned} A_k(2\omega) &:= \frac{E(e^{in_k\omega}\overline{P(z)}\Phi(z))}{E(P(z)\overline{P(z)}\Phi(z))}, \\ B_k(2\omega) &:= \frac{1}{E(P(\omega)\overline{P(z)}\Phi(z))}, \\ G_k(z) &:= \frac{1}{2^{d/2}} E(P(z)\overline{P(z)}\Phi(z))e^{in_k\omega} \\ &\quad - \frac{1}{2^{d/2}} E(e^{in_k\omega}\overline{P(z)}\Phi(z))P(z). \end{aligned}$$

Then  $G_k$  is orthogonal to  $V_0$  for all  $k = 1, \dots, 2^d$  and

$$V_1 = V_0 \bigoplus (G_1 + \dots + G_{2^d}).$$

Next we can show that integer translates of  $g_k, k = 1, \dots, 2^d$  are linearly dependent. However, under a certain condition on  $P$ , integer translates of  $g_k, k = 2, \dots, 2^d$  are linearly independent. Let us write  $P$  in its polyphase form, i. e.,

$$P(z) = \sum_{k=1}^{2^d} e^{in_k\omega} P_k(z^2), \quad (84)$$

where  $n_k, k = 1, \dots, 2^d$  are the multi-integers in the collection  $\{0, 1\}^d$  as defined above. We refer to [49] for a proof of the following theorem.

**Theorem 27** Suppose that  $\Phi(z^2) \neq 0$ . Suppose that there exists an integer  $k \in \{0, 1\}^d$  such that  $P_k(z^2) \neq 0$  for all  $z$  on the torus  $T^d$ . For simplicity, let us assume that  $P_1(z^2) \neq 0$  for all  $z$  on  $T^d$ . Then the integer translates of  $g_k, k = 2, \dots, 2^d$  form a Riesz basis for  $V_1 \ominus V_0$ .

The second step is to use a technique like the well-known Gram–Schmidt orthonormalization to construct  $\psi_{j,k}$  from  $g_{j,k}$  such that  $\psi_{j,k}$  are orthogonal among each other. It is a standard technique (cf. e. g., [39]).

We first choose  $\psi_2 = g_2$ . Let

$$\psi_3(x) = \sum_{m \in \mathbb{Z}^d} (c_{1,m}\psi_2(x-m) + c_{2,m}g_3(x-m))$$

for some coefficients  $c_{1,m}$  and  $c_{2,m}$ . To compute these coefficients, we write them in terms of Fourier transform

$$\begin{aligned} \widehat{\psi_3}(2\omega) &= C_1(z^2)\widehat{\psi_2}(2\omega) + C_2(z^2)\widehat{g_3}(2\omega) \\ &= (C_1(z^2)G_2(z) + C_2(z^2)G_3(z))\widehat{\phi}(\omega/2), \end{aligned}$$

where  $C_1$  and  $C_2$  are discrete Fourier transform of sequences  $c_{1,m}$ 's and  $c_{2,m}$ 's. For convenience, we let  $Q_2(z) = G_2(z)$  and

$$Q_3(z) = C_1(z^2)G_2(z) + C_2(z^2)G_3(z).$$

In order to have  $W_3 \perp W_2$ , the orthogonal condition in Theorem 26 implies that

$$\begin{aligned} C_1(z^2)E(G_2(z)\overline{G_2(z)}\Phi(z)) \\ + C_2(z^2)E(G_3(z)\overline{G_2(z)}\Phi(z)) = 0. \end{aligned} \quad (85)$$

By choosing

$$\begin{aligned} C_1(z^2) &= E(G_3(z)\overline{G_2(z)}\Phi(z)), \\ C_2(z^2) &= -E(G_2(z)\overline{G_2(z)}\Phi(z)), \end{aligned}$$

we know that the Eq. (85) holds and  $W_3$  is perpendicular to  $W_2$ . We continue this procedure above. To be more precise, let us show how to construct  $\psi_4$ . That is, let

$$\begin{aligned} \psi_4(x) &= \sum_{m \in \mathbb{Z}^d} (d_{1,m}\psi_2(x-m) + d_{2,m}\psi_3(x-m)) \\ &\quad + d_{3,m}g_4(x-m). \end{aligned}$$

In terms of Fourier transform, we have

$$\begin{aligned} \widehat{\psi_4}(2\omega) &= D_1(z^2)\widehat{\psi_2}(\omega) + D_2(z^2)\widehat{\psi_3}(\omega) + D_3(z^2)\widehat{g_4}(\omega) \\ &= (D_1(z^2)Q_2(z) + D_2(z^2)Q_3(z) \\ &\quad + D_3(z^2)G_4(z))\widehat{\phi}(\omega/2). \end{aligned}$$

In order to have  $W_4 \perp W_2$  and  $W_4 \perp W_3$ , we have the following two equations with three unknowns:

$$\begin{aligned} D_1(z^2)E(Q_2(z)\overline{Q_2(z)}\Phi(z)) \\ + D_3(z^2)E(G_4(z)\overline{Q_2(z)}\Phi(z)) = 0 \\ D_2(z^2)E(Q_3(z)\overline{Q_3(z)}\Phi(z)) \\ + D_3(z^2)E(G_4(z)\overline{Q_3(z)}\Phi(z)) = 0 \end{aligned} \quad (86)$$

which is an upper triangular homogeneous linear system. It can be solved easily. A solution may be given below. Let

$$\begin{aligned} D_1(z^2) &= E(Q_3(z)\overline{Q_3(z)}\Phi(z))E(G_4(z)\overline{Q_2(z)}\Phi(z)) \\ D_2(z^2) &= E(Q_2(z)\overline{Q_2(z)}\Phi(z))E(G_4(z)\overline{Q_3(z)}\Phi(z)) \\ D_3(z^2) &= -E(Q_2(z)\overline{Q_2(z)}\Phi(z))E(Q_3(z)\overline{Q_3(z)}\Phi(z)). \end{aligned}$$

With these Laurent polynomials  $D_1, D_2, D_3$ , the two equations in (86) are satisfied simultaneously. Thus we obtain the desired function  $\psi_4$ . Repeating the above constructive steps when  $d > 2$ , we find  $\psi_k, k = 2, \dots, 2^d$ . It is easy to see that  $\psi_k$ 's are compactly support when  $\phi$  are compactly supported. The above construction shows that the integer translates of  $\psi_k$  form a Riesz basis for  $W_k$  for  $k = 2, \dots, 2^d$  and  $W_k$ 's are mutually orthogonal. We have thus obtained the following

**Theorem 28** If a refinable function  $\phi$  generates an MRA for  $L_2(\mathbb{R}^d)$ . If  $\Phi(z) \neq 0$  for all  $z = e^{i\omega}$  and one of the polyphases of the mask  $P(\omega)$  of  $\phi$  is not zero for all  $e^{i\omega}$ . Then the functions  $\psi_{j,k}$  constructed above are prewavelets for  $L_2(\mathbb{R}^d)$  satisfying the conditions 1–4. in Definition 6.

Next we show how to use box splines to construct prewavelets in  $L_2(\mathbb{R}^d)$  since multivariate box splines are a very important class of refinable functions. Let us recall the definition of box splines. Let  $D$  be a set of nonzero vectors in  $\mathbb{R}^d$  (counting multiple of a same vector) which span  $\mathbb{R}^d$ . The box spline  $\phi_D$  associates with the direction set  $D$  is the function whose Fourier transform is defined by

$$\hat{\phi}_D(\omega) = \prod_{y \in D} \frac{1 - e^{-i\xi \cdot \omega}}{iy \cdot \omega}.$$

It is well-known that box spline  $\phi_D$  is a piecewise polynomial function of degree  $\leq \#D - d$ , where  $\#D$  denotes the cardinality of  $D$ . For more properties of box splines, see [7,11,60]. In particular, for  $d = 2$ ,  $e_1 = (1, 0)^T$ ,  $e_2 = (0, 1)^T$ , and

$$D = \underbrace{\{e_1, \dots, e_1\}}_{\ell}, \underbrace{\{e_2, \dots, e_2\}}_m, \underbrace{\{e_1 + e_2, \dots, e_1 + e_2\}}_n,$$

the box spline  $\phi_{\ell mn}$  based on such direction set  $D$  is called 3-direction box spline whose Fourier transform is

$$\hat{\phi}_{\ell mn}(\omega_1, \omega_2) = \left( \frac{1 - e^{-i\omega_1}}{i\omega_1} \right)^\ell \left( \frac{1 - e^{-i\omega_2}}{i\omega_2} \right)^m \left( \frac{1 - e^{-i(\omega_1 + \omega_2)}}{i(\omega_1 + \omega_2)} \right)^n.$$

It is well-known that box spline  $\phi_D$  generates a bona fide MRA of  $L_2(\mathbb{R}^d)$  (cf. [74]) when the direction set  $D$  is unimodular, i.e., the determinant of any  $d$  directions which span  $\mathbb{R}^d$  is 1 or  $-1$  (cf. [7]). The unimodularity also implies  $\Phi(\omega) > 0$ . Let  $\Phi_D$  be the Euler-Frobenius polynomial associated with  $\phi_D$  and  $P_D$  be the mask associated with  $\phi_D$ , i.e.,  $\widehat{\phi}_D(2\omega) = P_D(z)\widehat{\phi}_D(\omega)$ .

**Theorem 29** Consider the linear box spline in  $\mathbb{R}^d$ . That is, let

$$D = \{e_1, \dots, e_d, -(e_1 + \dots + e_d)\},$$

where  $e_i$  denotes the standard unit vector in  $\mathbb{R}^d$  which is 1 in the  $i$ th component while zero in the rest of the components for  $i = 1, \dots, d$ . Then  $\Phi_D(z) \neq 0$  for all  $z$  with  $|z| = 1$  and  $E(P_D(z)) = \frac{1}{2^d}$ .

*Proof* Since the  $D$  is unimodular, we have  $\Phi_D(z) \neq 0$  for all  $z$  with  $|z| = 1$ . Next it is easy to see that

$P_D(z) = \prod_{i=1}^d \left( \frac{1+z_i}{2} \right) \left( \frac{1+1/(z_1 \dots z_d)}{2} \right)$ . Then we can see that the even index term  $E(P_D(z))$  is only the constant term which is  $2/2^{d+1} = 1/2^d$ . This completes the proof.  $\square$

**Example 26** Consider  $\phi_{2,2,1}$ . Since  $P_{2,2,1}(z) = (1 + z_1)^2(1 + z_2)^2(1 + z_1 z_2)/32$ , it is easy to check that

$$E(P_{2,2,1}(z)) = \frac{1}{32} (5z_1^2 z_2^2 + z_1^2 + z_2^2 + 1).$$

Since

$$32|E(P_{2,2,1}(z))| = |5 + (z_1 z_2)^{-2} + (z_1)^{-2} + (z_2)^{-2}| > 5 - |(z_1 z_2)^{-2}| - |(z_1)^{-2}| - |(z_2)^{-2}| = 2,$$

we know that  $E(P_{2,2,1}(z)) \neq 0$  for  $z = (z_1, z_2)$  with  $|z_1| = |z_2| = 1$ .

**Example 27** Consider  $\phi_{2,2,2}$ . Similar to the examples above, we have

$$E(P_{2,2,2}(z)) = \frac{1}{64x^2 y^2} (10x^2 y^2 + x^2 + y^2 + 1 + y^4 x^2 + x^4 y^4 + x^4 y^2).$$

We can easily see that  $E(P_{2,2,2}(z)) \neq 0$ .  $\square$

**Example 28** We use box spline  $\tilde{B}_{1,1,1} = \phi_D$  based on  $D = \{e^1, e^2, -(e^1 + e^2)\}$  to construct compactly supported pre-wavelets in  $L_2(\mathbb{R}^2)$ . Note that our prewavelets have a larger support than those constructed in [44] and [39]. The purpose of this example is to show the detail of our constructive procedure. Clearly,

$$P(z) = \frac{1+z_1}{2} \frac{1+z_2}{2} \frac{1+1/(z_1 z_2)}{2}$$

and  $\Phi_D(z) = \frac{1}{2} + \frac{1}{12}(z_1 + z_2 + 1/z_1 + 1/z_2 + z_1 z_2 + 1/(z_1 z_2)) = \frac{1}{2} + \frac{1}{6}(\cos(\omega_1) + \cos(\omega_2) + \cos(\omega_1 + \omega_2)) \neq 0$  for any  $\omega_1$  and  $\omega_2$ . Using a computer algebra program Maple, we obtain the Laurent polynomials for  $G_1, \dots, G_4$  and  $Q_2, Q_3, Q_4$ . They are as follows:

$$768G_1(z_1, z_2) = \begin{bmatrix} z_1^{-3} \\ z_1^{-2} \\ z_1^{-1} \\ 1 \\ z_1 \\ z_1^2 \\ z_1^3 \end{bmatrix}^T \begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & 14 & -2 & 14 \\ -1 & -2 & -19 & -19 \\ -1 & 14 & -19 & 60 \\ 0 & -1 & -2 & -19 \\ 0 & 0 & -1 & 14 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} z_2^{-3} \\ z_2^{-2} \\ z_2^{-1} \\ 1 \\ z_2 \\ z_2^2 \\ z_2^3 \end{bmatrix},$$

$$768G_2(z_1, z_2) = \begin{bmatrix} z_1^{-1} \\ 1 \\ z_1 \\ z_1^2 \\ z_1^3 \end{bmatrix}^T \begin{bmatrix} -2 & 14 & -10 & 6 \\ -2 & -4 & -12 & -20 \\ 0 & 14 & -12 & 76 \\ 0 & 0 & -10 & -20 \\ 0 & 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} z_2^{-3} \\ z_2^{-2} \\ z_2^{-1} \\ 1 \\ z_2 \\ z_2^2 \\ z_2^3 \end{bmatrix},$$

$$768G_3(z_1, z_2) = \begin{bmatrix} z_1^{-3} \\ z_1^{-2} \\ z_1^{-1} \\ 1 \\ z_1 \\ z_1^2 \\ z_1^3 \end{bmatrix}^T \begin{bmatrix} -2 & -2 & 0 \\ 14 & -4 & 14 \\ -10 & -12 & -12 \\ 6 & -20 & 76 \\ 0 & -10 & -12 \\ 0 & 0 & 14 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_2^{-1} \\ 1 \\ z_2 \\ z_2^2 \\ z_2^3 \end{bmatrix},$$

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -10 & 0 \\ -20 & 6 \\ -12 & -10 \\ -4 & 14 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} z_2^{-1} \\ 1 \\ z_2 \\ z_2^2 \\ z_2^3 \end{bmatrix},$$

and

$$768G_4(z_1, z_2) = \begin{bmatrix} z_1^{-1} \\ 1 \\ z_1 \\ z_1^2 \\ z_1^3 \end{bmatrix}^T \begin{bmatrix} 6 & -10 & 14 \\ -10 & -20 & -12 \\ 14 & -12 & 76 \\ -2 & -4 & -12 \\ 0 & -2 & 14 \end{bmatrix} \begin{bmatrix} z_2^{-1} \\ 1 \\ z_2 \\ z_2^2 \\ z_2^3 \end{bmatrix}.$$

Since  $Q_2 = G_2$ , we now give  $Q_3$  as follows. Let

$$10616832Q_3 = [z_1^{-7}, \dots, z_1^{-1}, 1, z_1, \dots, z_1^{-7}]$$

$$Q[z_2^{-7}, \dots, z_2^{-1}, 1, z_2, \dots, z_2^{-9}]^T$$

with matrix  $Q$  being a of size  $15 \times 17$  defined by  $Q = [Q_1 Q_2]$  and  $Q_1$  of size  $15 \times 9$  and  $Q_2$  of size  $15 \times 8$ ,

where

$$Q_1 = \begin{bmatrix} -1 & 5 & -1 & 9 & 1 \\ 1 & -2 & -26 & -2 & -52 \\ -2 & 7 & 11 & 73 & 71 \\ 2 & -4 & -46 & 24 & -516 \\ -1 & -1 & 33 & 37 & 369 \\ 1 & -2 & -22 & 70 & -672 \\ 0 & -3 & 21 & -45 & 385 \\ 0 & 0 & -2 & 44 & -228 \\ 0 & 0 & 0 & -18 & 70 \\ 0 & 0 & 0 & 0 & -4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\begin{bmatrix} 3 & 1 & -1 & 0 \\ 2 & -22 & 2 & 3 \\ 63 & 85 & -17 & 27 \\ 144 & -628 & 168 & -158 \\ 271 & 773 & 15 & 480 \\ 714 & -3110 & 1402 & -2285 \\ -63 & 1713 & 291 & 2057 \\ 700 & -2462 & 2712 & -6520 \\ -174 & 746 & -260 & 2057 \\ 96 & -340 & 792 & -2285 \\ -28 & 74 & -150 & 480 \\ 0 & 2 & 52 & -158 \\ 0 & 0 & -14 & 27 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and

$$Q_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -14 & 0 & 0 & 0 \\ 52 & 2 & 0 & 0 \\ -150 & 74 & -28 & 0 \\ 792 & -340 & 96 & -4 \\ -260 & 746 & -174 & 70 \\ 2712 & -2462 & 700 & -228 \\ 291 & 1713 & -63 & 385 \\ 1402 & -3110 & 714 & -672 \\ 15 & 773 & 271 & 369 \\ 168 & -628 & 144 & -516 \\ 2 & -22 & 2 & -52 \\ -1 & 1 & 3 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -18 & 0 & 0 \\ 44 & -2 & 0 & 0 \\ -45 & 21 & -3 & 0 \\ 70 & -22 & -2 & 1 \\ 37 & 33 & -1 & -1 \\ 24 & -46 & -4 & 2 \\ -2 & -26 & -2 & 1 \\ 9 & -1 & 5 & -1 \end{bmatrix}.$$

The expression for  $Q_4$  involves a matrix of size about  $51 \times 51$ . Due to the space limit, we omit the details for  $Q_4$ .  $\square$

### Multivariate Tight Wavelet Frames and Bi-Frames

In this section, we generalize one of three constructive methods in Sect. “Tight Wavelet Frames” to give a construction of tight wavelet frames in the multivariate setting. We first introduce some notation.

Given a function  $\psi \in L_2(\mathbb{R}^d)$ , we let

$$\psi_{j,k}(y) = 2^{jd/2} \psi(2^j y - k), \forall j \in \mathbb{Z}, k \in \mathbb{Z}^d.$$

Let  $\Psi$  be a finite subset of  $L_2(\mathbb{R}^d)$  and

$$\Lambda(\Psi) := \{\psi_{j,k}; \psi \in \Psi, j \in \mathbb{Z}, k \in \mathbb{Z}^d\}$$

where  $\mathbb{Z}$  is the set of all integers.

**Definition 7** We say  $\Lambda(\Psi)$  is a frame if there exist two positive numbers  $A$  and  $B$  such that

$$A\|f\|_{L_2(\mathbb{R}^d)}^2 \leq \sum_{g \in \Lambda(\Psi)} |\langle f, g \rangle|^2 \leq B\|f\|_{L_2(\mathbb{R}^d)}^2$$

for all  $f \in L_2(\mathbb{R}^d)$ .  $\Lambda(\Psi)$  is a tight frame if it is a frame with  $A = B$ . In this case, after a renormalization of the  $g$ 's in  $\Psi$ , we have

$$\sum_{g \in \Lambda(\Psi)} |\langle f, g \rangle|^2 = \|f\|_{L_2(\mathbb{R}^d)}^2 \quad (87)$$

for all  $f \in L_2(\mathbb{R}^d)$ .

By polarization technique (as in Sect. “Tight Wavelet Frames”), the Eq. (87) implies that when  $\Lambda(\Psi)$  is a tight frame, any  $f \in L_2(\mathbb{R}^d)$  can be represented as

$$f = \sum_{g \in \Lambda(\Psi)} \langle f, g \rangle g.$$

We start with a compactly supported refinable function  $\phi \in L_2(\mathbb{R}^d)$  which generates a MRA of  $L_2(\mathbb{R}^d)$  under a standard dilation matrix  $2I_d$ , where  $I_d$  is the identity matrix in  $\mathbb{R}^d$ . Since  $\phi \in L_2(\mathbb{R}^d)$  is compactly supported and refinable,

$$\hat{\phi}(\omega) = P(\omega/2)\hat{\phi}(\omega/2)$$

where mask  $P(\omega)$  is a trigonometric polynomial. Suppose  $P$  satisfies

$$\sum_{j \in \{0,1\}^d \pi} |P(\omega + j)|^2 \leq 1. \quad (88)$$

Note that this condition is necessary, but may not be sufficient in the multivariate setting. We have to assume that there exist Laurent polynomials  $P_\ell$ ,  $\ell = 1, \dots, N$  such that

$$1 - \sum_{j \in \{0,1\}^d \pi} |P(\omega + j)|^2 = \sum_{k=1}^N |\tilde{P}_k(2\omega)|^2. \quad (89)$$

Here  $N$  is a nonnegative integer which is dependent on  $P$ . We shall show that for all multivariate box spline functions the condition (89) will be satisfied.

To construct tight wavelet frames we use the unitary extension principle (UEP) (cf. [76,77]). That is, we look for  $Q_i$  (trigonometric polynomial) such that

$$\begin{aligned} P(\omega) \overline{P(\omega + \ell)} + \sum_{i=0}^r Q_i(\omega) \overline{Q_i(\omega + \ell)} \\ = \begin{cases} 1 & \text{if } \ell = 0, \\ 0, & \ell \in \{0,1\}^d \pi \setminus \{0\}. \end{cases} \end{aligned} \quad (90)$$

With these  $Q_i$ 's we can define wavelet frame generators  $\psi^{(i)}$ , in terms of their Fourier transforms, by

$$\hat{\psi}^{(i)}(\omega) = Q_i(\omega/2)\hat{\phi}(\omega/2), \quad i = 1, \dots, r. \quad (91)$$

Then, if  $\phi$  is continuous and Lip  $\alpha$ , with  $\alpha > 0$ , and the UEP is satisfied, the family  $\Psi = \{\psi^{(i)}, i = 1, \dots, r\}$  generates a tight frame, i.e.,  $\Lambda(\Psi)$  is a tight wavelet frame. This result can be proved by using the same proof of Theorem 12 in Sect. “Tight Wavelet Frames”. (See [20] and [25] for different proofs.)

For convenience, we rewrite (90) in an equivalent matrix form as follows:

**Lemma 26** Let  $\mathcal{P} = (P(\omega + \ell); \ell \in \{0,1\}^d \pi)^T$  be a vector of size  $2^d \times 1$  and  $\mathcal{Q} = (Q_i(\omega + \ell); \ell \in \{0,1\}^d \pi, i = 1, \dots, r)$  be a matrix of size  $2^d \times r$ . Then (90) is equivalent to

$$\mathcal{Q}\mathcal{Q}^* = I_{2^d} - \mathcal{P}\mathcal{P}^*, \quad (92)$$

where  $P^*$  denotes the complex conjugate transpose of the column vector  $P$ .

*Proof* This can be verified directly.  $\square$

For example, when  $d = 2$ ,  $r = 4$  and  $\omega = (\xi, \eta)$ , we have

$$Q = \begin{bmatrix} Q_1(\xi, \eta) & Q_1(\xi + \pi, \eta) \\ Q_2(\xi, \eta) & Q_2(\xi + \pi, \eta) \\ Q_3(\xi, \eta) & Q_3(\xi + \pi, \eta) \\ Q_4(\xi, \eta) & Q_4(\xi + \pi, \eta) \\ Q_1(\xi, \eta + \pi) & Q_1(\xi + \pi, \eta + \pi) \\ Q_2(\xi, \eta + \pi) & Q_2(\xi + \pi, \eta + \pi) \\ Q_3(\xi, \eta + \pi) & Q_3(\xi + \pi, \eta + \pi) \\ Q_4(\xi, \eta + \pi) & Q_4(\xi + \pi, \eta + \pi) \end{bmatrix}^T,$$

$P = (P(\xi, \eta), P(\xi + \pi, \eta), P(\xi, \eta + \pi), P(\xi + \pi, \eta + \pi))^T$ , and

$$QQ^* = I_{2^d} - PP^*. \quad (93)$$

Our construction of tight wavelet frames is mainly based on the matrix form (92).

In general, for  $P$  satisfying (88) we write  $P$  in its polyphase form (cf. (84) with a normalized constant) and let  $\widehat{P} = (P_m(2\omega); m \in \{0, 1\}^d)^T = \mathcal{M}^*P$ , where  $\mathcal{M}$  is the polyphase matrix

$$\mathcal{M} = 2^{-d/2} (e^{im \cdot (\omega + \ell)})_{\substack{\ell \in \{0, 1\}^d \pi \\ m \in \{0, 1\}^d}} \quad (94)$$

which is unitary and  $P = (P(\omega + \ell); \ell \in \{0, 1\}^d \pi)^T$ . Here we have abused notation by writing  $P(\omega)$  in terms of  $P(z)$  with  $z = e^{i\omega}$ . Then (88) is equivalent to

$$\widehat{P}^* \widehat{P} = \sum_{m \in \{0, 1\}^d} |\widehat{P}_m(2\omega)|^2 \leq 1. \quad (95)$$

**Theorem 30** Suppose that  $P$  satisfies the condition (88). Suppose that there exist Laurent polynomials  $\widetilde{P}_1, \dots, \widetilde{P}_N$  such that

$$\sum_{m \in \{0, 1\}^d} |P_m(\omega)|^2 + \sum_{i=1}^N |\widetilde{P}_i(\omega)|^2 = 1. \quad (96)$$

Then there exist  $2^d + N$  compactly supported tight frame generators with wavelet masks  $Q_m$ ,  $m = 1, \dots, 2^d + N$ , such that  $P, Q_m$ ,  $m = 1, \dots, 2^d + N$ , satisfy (90).

*Proof* We define the combined column vector  $\widetilde{P} = (P_m(2\omega); m \in \{0, 1\}^d, \widetilde{P}_i(2\omega); 1 \leq i \leq N)^T$  of size  $(2^d + N)$  and the matrix

$$\widetilde{Q} := I_{(2^d + N)} - \widetilde{P} \widetilde{P}^*.$$

Note that all entries of  $\widetilde{P}$  and  $\widetilde{Q}$  are  $\pi$ -periodic. Identity (96) implies that  $\widetilde{Q} \widetilde{Q}^* = \widetilde{Q}$ , and this gives

$$\widetilde{P} \widetilde{P}^* + \widetilde{Q} \widetilde{Q}^* = I_{(2^d + N)}.$$

Restricting to the first principle  $2^d \times 2^d$  blocks in the above matrices, we have

$$\widehat{P} \widehat{P}^* + \widehat{Q} \widehat{Q}^* = I_{2^d}, \quad (97)$$

where  $\widehat{P} = \mathcal{M}^*P$  was already defined above and  $\widehat{Q}$  denotes the first  $2^d \times (2^d + N)$  block matrix of  $\widetilde{Q}$ . By (94), we have  $P = \mathcal{M} \widehat{P}$ , and (97) yields

$$PP^* + \mathcal{M} \widehat{Q} (\mathcal{M} \widehat{Q})^* = I_{2^d},$$

which is (93). Thus we let

$$Q = \mathcal{M} \widehat{Q}.$$

Then the first row  $[Q_1, \dots, Q_{2^d + N}]$  of  $Q$  gives the desired trigonometric functions for compactly supported tight wavelet frame generators. The form  $Q = [Q_i(\omega + \ell)]$  is inherited from  $\mathcal{M}$ , since the entries of  $\widehat{Q}$  are  $\pi$ -periodic. This completes the proof.  $\square$

In general, we do not know if (96) holds for any given mask  $P$ . The problem is related to Hilbert's 17th problem. For any multivariate nonnegative polynomial, it is not known that it can be written as a sum of square of finitely many polynomials (cf. [30,80]). However, we have the following

**Theorem 31** Let  $P(\omega) = \sum_{k \in \mathbb{Z}^d} c_k z^k$  be a Laurant polynomial, with  $N$  nonzero coefficients  $c_k$ . Suppose that all  $c_k$  are nonnegative. Furthermore, writing

$$P(\omega) = \sum_{k=1}^{2^d} e^{i\pi n_k} P_k(2\omega)$$

in its polyphase form with  $\{n_k, k = 1, \dots, 2^d\} = \{0, 1\}^d$ , suppose that  $P_k(0) = 1/2^d$ ,  $k = 1, \dots, 2^d$ . Then there exist at most  $r = N^2$  polynomials  $Q_j$  such that (90) holds.

A proof of this result can be found in [10]. In particular, for a box spline function  $\phi$ , its mask polynomial satisfies the condition in Theorem 31. Thus we can always use box spline function to construct tight wavelet frames.

We shall use the constructive scheme above to find compactly supported tight wavelet frames based on multivariate box splines, in particular, bivariate box splines on three and four directional meshes.

Let us recall the definition of box spline  $\phi_D$  from Sect. "Multivariate Prewavelets", where  $D$  is a set of non zero vectors in  $\mathbb{R}^d$  (allowing multiples of the same vector)

which span  $\mathbb{R}^d$ . It is well-known that  $\phi_D$  is refinable and its Fourier transform satisfies  $\widehat{\phi}_D(\omega) = P_D(\frac{\omega}{2})\widehat{\phi}_D(\frac{\omega}{2})$  that

$$P_D(\omega) = \prod_{\xi \in D} \frac{1 + e^{-i\xi \cdot \omega}}{2}.$$

Thus  $P_D$  is a Laurent polynomial and  $|P_D(\omega)|^2 = \prod_{\xi \in D} (\cos \frac{\xi \cdot \omega}{2})^2$ . Then we have the following result.

**Lemma 27** Suppose that a given direction set  $D \subset \mathbb{Z}^d$  contains all of the standard unit vectors  $e_i$  of  $\mathbb{R}^d$ ,  $i = 1, \dots, d$ . Then  $P_D$  satisfies (88).

*Proof* Since  $|P_D(\omega)|^2 \leq \prod_{i=1}^d \cos^2 \frac{\omega_i}{2}$ , with  $\omega = (\omega_1, \dots, \omega_s)^T \in \mathbb{R}^d$ , we have

$$\sum_{\ell \in \{0,1\}^d \pi} |P_D(\omega + \ell)|^2 \leq \prod_{i=1}^d \left( \cos^2 \frac{\omega_i}{2} + \sin^2 \frac{\omega_i}{2} \right) = 1.$$

This completes the proof.  $\square$

We now give some examples that the mask polynomial  $P_D$  associated with bivariate box splines on three and four direction meshes satisfy (89).

**Example 29** Consider a three directional box spline  $\phi_{1,1,1}$ . It is easy to see that

$$\begin{aligned} 1 - \sum_{\ell \in \{0,1\}^2 \pi} |P_{1,1,1}(\omega + \ell)|^2 \\ = \frac{3}{8} - \frac{1}{8} \cos(2\omega_1) - \frac{1}{8} \cos(2\omega_2) - \frac{1}{8} \cos(2\omega_1 + 2\omega_2). \end{aligned}$$

Thus, we let

$$\begin{aligned} \widetilde{P}_1(\omega) &= \frac{\sqrt{6}}{8}(1 - e^{i\omega_1}), \quad \text{and} \\ \widetilde{P}_2(\omega) &= \frac{\sqrt{2}}{8}(2 - e^{i\omega_2} - e^{i(\omega_1 + \omega_2)}). \end{aligned}$$

Clearly, we have

$$\sum_{\ell \in \{0,1\}^2 \pi} |P_{1,1,1}(\omega + \ell)|^2 + \sum_{i=1}^2 |\widetilde{P}_i(2\omega)|^2 = 1.$$

Thus, one can apply the constructive steps in the proof of Theorem 30 to get 6 tight frame masks  $Q_i$ ,  $i = 1, \dots, 6$ .  $\square$

**Example 30** Consider box spline  $\phi_{2,2,1}$ . We find that

$$\begin{aligned} 1 - \sum_{\ell \in \{0,1\}^2 \pi} |P_{2,2,1}(\omega + \ell)|^2 \\ = \frac{19}{32} - \frac{7}{32} \cos(2\omega_1) - \frac{7}{32} \cos(2\omega_2) \\ - \frac{1}{64} \cos(2\omega_1 - 2\omega_2) - \frac{9}{64} \cos(2\omega_1 + 2\omega_2). \end{aligned}$$

Let

$$\begin{aligned} \widetilde{P}_1(\omega) &= \frac{\sqrt{21}}{12} - \frac{\sqrt{102} + 2\sqrt{21}}{48} e^{i\omega_1} \\ &\quad + \frac{\sqrt{102} - 2\sqrt{21}}{48} e^{i\omega_2} \\ \widetilde{P}_2(\omega) &= -\frac{\sqrt{42} + 2\sqrt{51}}{48} + \frac{\sqrt{42}}{24} e^{i\omega_2} \\ &\quad - \frac{\sqrt{42} - 2\sqrt{51}}{48} e^{i(\omega_1 + \omega_2)}. \end{aligned}$$

It is easy to check that

$$\sum_{\ell \in \{0,1\}^2 \pi} |P_{2,2,1}(\omega + \ell)|^2 + \sum_{i=1}^2 |\widetilde{P}_i(2\omega)|^2 = 1.$$

Hence, the constructive steps in the proof of Theorem 30 yield 6 tight frame masks and thus, 6 tight frame generators.  $\square$

**Example 31** For box spline  $\phi_{1,1,1,1}$ , we have

$$\begin{aligned} 1 - \sum_{\ell \in \{0,1\}^2 \pi} |P_{1,1,1,1}(\omega + \ell)|^2 \\ = \frac{5}{8} - \frac{1}{8} (e^{i2\omega_1} + e^{-i2\omega_1}) - \frac{1}{8} (e^{i2\omega_2} + e^{-i2\omega_2}) \\ - \frac{1}{32} (e^{2i(\omega_1 + \omega_2)} + e^{-2i(\omega_1 + \omega_2)}) \\ - \frac{1}{32} (e^{2i(\omega_1 - \omega_2)} + e^{-2i(\omega_1 - \omega_2)}) = \sum_{i=1}^2 |\widetilde{P}_i(2\omega)|^2, \end{aligned}$$

where  $\widetilde{P}_1(\omega) = \frac{\sqrt{6}}{8}(1 - e^{i(\omega_1 - \omega_2)})$ , and

$$\widetilde{P}_2(\omega) = -\frac{1}{4} + \frac{\sqrt{6}}{8} + \frac{1}{4}(e^{i\omega_1} + e^{i\omega_2}) - \frac{2 + \sqrt{6}}{8} e^{i(\omega_1 + \omega_2)}.$$

Hence, the constructive steps in the proof of Theorem 30 yield 6 tight frame masks and hence, 6 tight frame generators.  $\square$

In the rest of this section, we construct compactly supported bi-frames. For refinable functions  $\phi$  and  $\phi^{\text{dual}}$ , let  $P$  and  $P^{\text{dual}}$  denote the symbols of the respective refinement masks. We use the superscript *dual* in order to point to duality of the respective frames. Moreover, let  $\psi_j$  and  $\psi_j^{\text{dual}}$  be functions associated with  $\phi$  and  $\phi^{\text{dual}}$  defined by

$$\begin{aligned} \hat{\psi}_j(\omega) &= Q_j(\omega/2)\hat{\phi}(\omega/2) \quad \text{and} \\ \widehat{\psi_j^{\text{dual}}}(\omega) &= Q_j^{\text{dual}}(\omega/2)\widehat{\phi^{\text{dual}}}(\omega/2), \end{aligned} \tag{98}$$

where  $Q_j$ ,  $Q_j^{\text{dual}}$ ,  $j = 1, \dots, r$ , are two families of masks. Let  $\Psi = \{\psi_j, j = 1, \dots, r$  and  $\Psi^{\text{dual}} = \{\psi_j^{\text{dual}}, j = 1,$

$\dots, r\}$  be the corresponding families of framelets and their shifts and dilates are collected into the following two sets:

$$\begin{aligned}\Lambda(\Psi) &:= \{2^{jd/2}\psi_i(2^jx - k); j \in \mathbb{Z}, k \in \mathbb{Z}^d, \\ &\quad i = 1, \dots, r\}, \\ \Lambda(\Psi^{\text{dual}}) &:= \{2^{jd/2}\psi_i^{\text{dual}}(2^jx - k); j \in \mathbb{Z}, k \in \mathbb{Z}^d, \\ &\quad i = 1, \dots, r\}.\end{aligned}\quad (99)$$

Recall that a family of functions  $\{\phi_j, j \in J\}$  is a Bessel family if

$$\left\| \sum_{j \in J} c_j \phi_j \right\|_2^2 \leq B \sum_{j \in J} |c_j|^2$$

for any coefficients  $c_j$  (see [17]).

**Definition 8** The two families  $\Lambda(\Psi)$  and  $\Lambda(\Psi^{\text{dual}})$  are called bi-frames if they are Bessel families, and the duality relation holds for all  $f, g \in L_2(\mathbb{R}^d)$

$$\langle f, g \rangle = \sum_{i=1}^n \sum_{\substack{j \in \mathbb{Z} \\ k \in \mathbb{Z}^d}} \langle f, \psi_{i;j,k} \rangle \left\langle \psi_{i;j,k}^{\text{dual}}, g \right\rangle. \quad (100)$$

The functions  $\psi_i$  and  $\psi_i^{\text{dual}}$  are called bi-framelets or bi-frame generators.

It follows that

$$f = \sum_{i=1}^n \sum_{\substack{j \in \mathbb{Z} \\ k \in \mathbb{Z}^d}} \langle f, \psi_{i;j,k} \rangle \psi_{i;j,k}^{\text{dual}} \quad (101)$$

and

$$f = \sum_{i=1}^n \sum_{\substack{j \in \mathbb{Z} \\ k \in \mathbb{Z}^d}} \left\langle \psi_{i;j,k}^{\text{dual}}, f \right\rangle \psi_{i;j,k} \quad (102)$$

weakly for all  $f \in L_2(\mathbb{R}^2)$ .

To construct the bi-frames, we need the following theorem (cf. Proposition 5.2 in [25]).

**Theorem 32** Suppose that  $\phi$  and  $\phi^{\text{dual}}$  are compactly supported refinable functions. Suppose that there are  $Q_i, Q_i^{\text{dual}}, i = 1, \dots, r$ , satisfying

$$P(\omega) \overline{P^{\text{dual}}(\omega + \ell)} + \sum_{i=0}^r Q_i(\omega) \overline{Q_i^{\text{dual}}(\omega + \ell)} = \delta_\ell \quad (103)$$

for  $\ell \in \{0, 1\}^d \pi$ . Suppose that  $Q_i(\omega)$  and  $Q_i^{\text{dual}}(\omega)$  have a zero at  $\omega = 0$ . Let  $\psi_i$  and  $\psi_i^{\text{dual}}$  be the functions defined

by their Fourier transform in (98). Then the two families  $\Lambda(\Psi)$  and  $\Lambda(\Psi^{\text{dual}})$  are bi-frames.

It is easy to see that the relations in (103) can be recast as

**Lemma 28** Let  $P = (P(\omega + \ell); \ell \in \{0, 1\}^d \pi)$  be a vector of size  $2^d \times 1$ ,  $Q = (Q_i(\omega + \ell); \ell \in \{0, 1\}^d \pi, i = 1, \dots, r)$  be a matrix of size  $2^d \times r$ , and  $P^{\text{dual}}, Q^{\text{dual}}$  be given analogously. Then (103) is equivalent to

$$Q(Q^{\text{dual}})^* = I_r - P(P^{\text{dual}})^*. \quad (104)$$

*Proof* This can be verified directly.  $\square$

We will construct compactly supported bi-frames for those masks  $P$  and  $P^{\text{dual}}$  which satisfy

$$\sum_{\ell \in \{0, 1\}^d \pi} P(\omega + \ell) \overline{P^{\text{dual}}(\omega + \ell)} = 1 \quad (105)$$

and  $P(0) = 1 = P^{\text{dual}}(0)$ . Let  $P$  and  $P^{\text{dual}}$  be given as in Lemma 28. Recall the unitary matrix  $\mathcal{M}$  defined as before. Then we have

**Theorem 33** Define

$$Q := (Q_i(\omega + \ell))_{\substack{\ell \in \{0, 1\}^d \pi \\ i=1, \dots, 2^d}} = (I_{2^d \times 2^d} - P(P^{\text{dual}})^*) \mathcal{M}$$

and

$$Q^{\text{dual}} := (Q_i^{\text{dual}}(\omega + \ell))_{\substack{\ell \in \{0, 1\}^d \pi \\ i=1, \dots, 2^d}} = (I_{2^d \times 2^d} - P^{\text{dual}} P^*) \mathcal{M}.$$

Then  $P, P^{\text{dual}}, Q$ , and  $Q^{\text{dual}}$  satisfy (104). Let  $\psi_i$  and  $\psi_i^{\text{dual}}$  be defined by (98), with these  $Q_i$ 's and  $Q_i^{\text{dual}}$ 's. Then  $\{\psi_i, i = 1, \dots, 2^d\}$  and  $\{\psi_i^{\text{dual}}, i = 1, \dots, 2^d\}$  are bi-framelets.

*Proof* It is trivial to verify that

$$Q(Q^{\text{dual}})^* = I_{2^d \times 2^d} - P(P^{\text{dual}})^*$$

which is (104) with  $S(\omega) = 1$ . Since both  $\mathcal{M}$  and  $P$  have the desired form, and since  $(P^{\text{dual}})^* \mathcal{M}$  is a row vector whose entries are  $\pi$  periodic, the matrix  $Q$  has the desired form as well. Analogous statements hold for  $Q^{\text{dual}}$ .

Next we need to verify the vanishing moment conditions for  $Q_i$  and  $Q_i^{\text{dual}}$ . Let  $(\hat{P}_m(2\omega); m \in \{0, 1\}^d) = \mathcal{M}^* P$  be the polyphase components of  $P$ . Then

$$Q_m^{\text{dual}}(\omega) = 2^{-d/2} e^{im \cdot \omega} - P^{\text{dual}}(\omega) \overline{\hat{P}_m(2\omega)}$$

and  $\hat{P}_m(0) = 2^{-d/2}$ . Note that  $P^{\text{dual}}(0) = 1$ . Therefore,  $Q_m^{\text{dual}}(0) = 0$  for  $m \in \{0, 1\}^d$ . Analogous statements show

that  $Q_m(0) = 0$  for  $m \in \{0, 1\}^d$ . Using Theorem 32, we conclude that  $\psi_i$  and  $\psi_i^{\text{dual}}$  defined above, using these  $Q_m$ 's and  $Q_m^{\text{dual}}$ 's, are bi-framelets. This completes the proof.  $\square$

We have the following example of bi-frame generators based on bivariate box splines.

**Example 32** For the mask  $P_{\ell, m, n}$ , associated with the bivariate box spline  $B_{\ell, m, n}$  on the three direction mesh, many dual masks  $P_{\ell, m, n}^{\text{dual}}$  were given in Sect. “Biorthogonal Box Spline Wavelets” satisfying (105) with  $d = 2$ . Then the formulae for  $Q$  and  $Q^{\text{dual}}$  given in Theorem 33 provide an explicit representation of bi-framelets or bi-frame generators.  $\square$

Next we consider a general refinable function  $\phi$ . Let  $P$  be the mask associated with  $\phi$ . Note that  $P(0) = 1$ . Assume that  $P(\ell) = 0$  for  $\ell \in \{0, 1\}^d \setminus \{0\}$ . To ensure (105) for any given mask  $P$ , we may use the celebrated Hilbert Nullstellensatz. Indeed, we let  $P_m(2\omega)$  be the polyphase components of  $P$ , i. e.,

$$(P_m(2\omega); m \in \{0, 1\}^d) = \mathcal{M}^*(P(\omega + \ell); \ell \in \{0, 1\}^d).$$

Similarly, for the dual mask  $P_{\text{dual}}$ , let  $P_m^{\text{dual}}(2\omega)$  be the polyphase components of  $P_{\text{dual}}$ . Then (105) is equivalent to

$$\sum_{m \in \{0, 1\}^d} P_m(\omega) \overline{P_m^{\text{dual}}(\omega)} = 1.$$

By the Hilbert Nullstellensatz, we have

**Lemma 29** Let  $P$  be the mask of a refinable function  $\phi$ . Write  $\hat{P}_m(z) := P_m(\omega)$  in terms of  $z = e^{i\omega} := (e^{i\omega_1}, \dots, e^{i\omega_d}) \in \mathbb{C}^d$  for  $m \in \{0, 1\}^d$ . If the Laurent polynomials  $\hat{P}_m$  have no common zero in  $(\mathbb{C} \setminus \{0\})^d$ , then there exist Laurent polynomials  $\hat{Q}_m(z)$  such that

$$\sum_{m \in \{0, 1\}^d} \hat{P}_m(z) \hat{Q}_m(z) = 1. \quad (106)$$

Thus, we let  $P^{\text{dual}}(\omega) = 2^{-d/2} \sum_{m \in \{0, 1\}^d} e^{im \cdot \omega} \hat{Q}_m(e^{i2\omega})$ . Then  $P$  and  $P^{\text{dual}}$  satisfy (105). In order to apply our Theorem 33, we only need to make sure that  $P^{\text{dual}}(0) = 1$ . Using the fact  $P(0) = 1$  and the assumption  $P(\ell) = 0$  for  $\ell \in \{0, 1\}^d \setminus \{0\}$ , we conclude from (105) that  $P^{\text{dual}}(0) = 1$ . Hence, we obtain the following

**Theorem 34** Given a mask  $P$ , suppose that  $P(0) = 1$  and  $P(\ell) = 0$  for  $\ell \in \{0, 1\}^d \setminus \{0\}$ . Let  $P_m(\omega)$ ,  $m \in \{0, 1\}^d$ , be the polyphase components of  $P$ . Writing  $\hat{P}_m(z) := P_m(\omega)$  in terms of  $z = e^{i\omega}$ , suppose that the Laurent polynomials  $\hat{P}_m$  have no common zero in  $z \in \mathbb{C}^d \setminus \{0\}$ . Then

there exists a pair of bi-frames  $\{\psi_i; i = 1, \dots, 2^d\}$  and  $\{\psi_i^{\text{dual}}; i = 1, \dots, 2^d\}$  associated with  $P$ .

The tight wavelet frames and bi-frames constructed above have one order of vanishing moment. In order to increase the order of vanishing moment, we have to introduce vanishing moment recovery function  $S$  or use *Oblique Extension Principle*. We refer the reader to [20] and [25] for the details.

### Spherical Tight Wavelet Frames

Let  $S \in \mathbb{R}^3$  be the unit spherical surface and  $L_2(S)$  be the space of all square integrable functions over  $S$ . For any function  $F(x, y, z) \in L_2(S)$  with  $|x|^2 + |y|^2 + |z|^2 = 1$ , we can use a standard transform  $x = \cos(\theta) \cos(\phi)$ ,  $y = \cos(\theta) \sin(\phi)$ , and  $z = \sin(\theta)$  to convert  $F$  into a function  $f$  over  $[-\pi/2, \pi/2] \times [0, 2\pi]$  by

$$f(\theta, \phi) = F(x, y, z).$$

Note that  $f(\theta, \phi)$  is not an ordinary function over rectangular domain  $[-\pi/2, \pi/2] \times [0, 2\pi]$ . When  $F$  is continuous at the north and south poles, we have

$$f(\pm\pi/2, \phi) = F(0, 0, \pm 1), \quad \forall \phi \in [0, 2\pi]. \quad (107)$$

To have  $C^1$  continuity or continuous tangent plane at the both poles, we have

$$\begin{aligned} \frac{\partial f}{\partial \theta}(\pm\pi/2, \phi) &= -F_1(0, 0, \pm 1) \cos(\phi) - F_2(0, 0, \pm 1) \\ &\quad \sin(\phi), \quad \forall \phi \in [0, 2\pi]. \end{aligned} \quad (108)$$

The conditions (107) and (108) can be found in [26, 66, 85].

Furthermore, using the standard transform, we have

$$\int_S |F(x, y, z)|^2 dS = \int_{-\pi/2}^{\pi/2} \int_0^{2\pi} |f(\theta, \phi)|^2 \cos(\theta) d\phi d\theta. \quad (109)$$

Thus, we consider a special  $L_2$  space over rectangular domain  $[-\pi/2, \pi/2] \times [0, 2\pi]$ . Let  $L_2^*([-\pi/2, \pi/2] \times [0, 2\pi])$  be the space of all bivariate functions  $f$  on  $[-\pi/2, \pi/2] \times [0, 2\pi]$  such that the right-hand side of the Eq. (109) is finite.

To build a multiresolution approximation of  $L_2^*([-\pi/2, \pi/2] \times [0, 2\pi])$ , we may use tensor product of two sets of univariate refinable functions. Let  $V_0 = \text{span}\{B_1(\theta), \dots, B_m(\theta)\}$  and  $\hat{V}_0 = \text{span}\{T_1(\phi), \dots,$



$T_n(\phi)\}$  be two basic spaces. Consider functions in the following form:

$$f(\theta, \phi) = \sum_{i=1}^m \sum_{j=1}^n c_{ij} B_i(\theta) T_j(\phi)$$

for coefficients  $c_{ij}$ . Assume that the basis functions  $B_i$  and  $T_j$  are locally supported, that is,  $B_i$  does not contain a neighborhood  $[-\pi/2, -\pi/2 + \delta)$  except for  $i = 1$  and  $B_i$  does not contain  $(\pi/2 - \delta, \pi/2]$  except for  $i = m$ , where  $\delta > 0$  is sufficiently small. In this case, we have

$$\begin{aligned} f(-\pi/2, \phi) &= \sum_{j=1}^n c_{1j} B_1(-\pi/2) T_j(\phi) \\ f(\pi/2, \phi) &= \sum_{j=1}^n c_{mj} B_m(\pi/2) T_j(\phi). \end{aligned}$$

By (107), we have

$$\begin{aligned} F(0, 0, -1) &= \sum_{j=1}^n c_{1j} B_1(-\pi/2) T_j(\phi) \\ F(0, 0, 1) &= \sum_{j=1}^n c_{mj} B_m(\pi/2) T_j(\phi) \end{aligned}$$

Similarly, by using (108), we have

$$\begin{aligned} -F_1(0, 0, -1) \cos(\phi) - F_2(0, 0, -1) \sin(\phi) \\ = \sum_{j=1}^n c_{1j} B'_1(-\pi/2) T_j(\phi) \\ -F_1(0, 0, 1) \cos(\phi) - F_2(0, 0, 1) \sin(\phi) \\ = \sum_{j=1}^n c_{mj} B'_m(\pi/2) T_j(\phi). \end{aligned}$$

Therefore, the space  $\text{span}\{T_j, j = 1, \dots, n\}$  has to be able to reproduce 1 and both  $\cos(\phi)$  and  $\sin(\phi)$ . Fortunately, trigonometric splines of even degrees have such a property (based on trigonometric Marsden's identity).

This is why in [66], a tensor of  $C^1$  quadratic B-splines and  $C^1$  trigonometric splines are used to build a multiresolution approximation of  $L_2(S)$  and to construct prewavelets. To have a higher smoothness at the both north and south poles, one has to use trigonometric splines of higher order (cf. [53]).

Trigonometric splines were first studied in [82] and they are much like ordinary B-splines. More literature on trigonometric splines can be found in [67,68,84]. Let us give a brief explanation of trigonometric B-splines. Let

$$s(x) = \sin(x), \quad c(x) = \cos(x).$$

Let  $\mathcal{T}_d$  be the space of all trigonometric polynomials of degree  $\leq d$ . That is,  $\mathcal{T}_d$  is a collection of functions of form

$$f(x) = \begin{cases} \frac{a_0}{2} + \sum_{j=1}^m (a_{2j} c(2jx) + b_{2j} s(2jx)), & \text{if } d = 2m \\ \sum_{j=1}^m (a_{2j-1} c((2j-1)x) + b_{2j-1} s((2j-1)x)), & \text{if } d = 2m-1. \end{cases}$$

It is easy to see that  $\mathcal{T}_d$  is a linear vector space of dimension  $d+1$ . It is easy to verify that functions  $c(2jx), s(2jx), j = 0, \dots, m$  are orthogonal with respect to  $L_2$  inner product on interval  $[0, 2\pi]$  when  $d = 2m$  is even. Similar for the case when  $d = 2m-1$  is odd. Any function  $f \in \mathcal{T}_d$  satisfies  $f(x + 2\pi) = (-1)^d f(x)$  for all  $x$  and  $d$ . That is,  $f$  is periodic on  $[0, 2\pi]$  if  $d$  is even.

Trigonometric splines of degree  $d$  are piecewise trigonometric functions, which each piece belongs to the space  $\mathcal{T}_d$  of trigonometric polynomials of degree  $\leq d$ . For simplicity, we consider  $[0, 2\pi]$ . Suppose that  $\mathbf{t} = (t_j, j = 0, \dots, n+1)$  is a knot sequence with length at least  $n \geq d+2$  such that  $0 < t_{j+1} - t_j < \pi$  for all possible  $j$ . Starting with

$$T_{0,j}(x) = \begin{cases} 1, & \text{if } t_j \leq x < t_{j+1} \\ 0, & \text{otherwise,} \end{cases} \quad (110)$$

for  $j = 0, \dots, n$ , and for  $k = 1, 2, \dots, d$ , we recursively define

$$\begin{aligned} T_{k,j}(x) &= \frac{s(x - t_j)}{s(t_{j+k} - t_j)} T_{k-1,j}(x) \\ &\quad + \frac{s(t_{j+1+k} - x)}{s(t_{j+1+k} - t_{j+1})} T_{k-1,j+1}(x) \end{aligned} \quad (111)$$

for  $j = 0, \dots, n-k$ , where terms with zero denominators are defined to be zero.  $T_{k,j}$  is called the  $j$ th trigonometric B-spline of degree  $k$ . Repeating (111) we have

**Lemma 30** A trigonometric B-spline  $T_{k,j}$  is a piecewise trigonometric polynomial. In fact,

$$T_{k,j}(x) = \sum_{i=j}^{j+k} P_{i,j,k}(x) T_{0,i}(x),$$

where  $P_{i,j,k}(x)$  are trigonometric polynomials  $\in \mathcal{T}_k$ .

More properties of trigonometric B-splines can be found in the references mentioned above. Especially, we can show that  $T_{k,j}$  is refinable. That is, after uniformly refining the knot sequence  $\mathbf{t}$ , we can define trigonometric B-spline  $T_{k,j}^{h/2}$  as above and show that  $T_{k,j}$  is a finitely linear combination of  $T_{k,j'}^{h/2}$ .

We are now ready to present a method to find a tight wavelet frame for  $L_2(S)$ . We begin with

**Definition 9** The family  $\{\Psi_k\}_{k \in \mathbb{Z}_+}$  of function in  $L_2(S)$  is a (MRA) tight wavelet frame for  $L^2(S)$  if

$$\|F\|^2 = \sum_{j \in \mathbb{Z}_+} |\langle F, \Psi_j \rangle|^2, \quad \forall f \in L^2(S).$$

Using the polarization technique as in Sect. “Tight Wavelet Frames” we have

$$F(x, y, z) = \sum_{k \in \mathbb{Z}_+} \langle f, \Psi_j \rangle \Psi_j(x, y, z)$$

for all  $F \in L_2(S)$ .

By (109) it is equivalent to build a tight wavelet frame for  $L_2^*([-\pi/2, \pi/2] \times [0, 2\pi])$ . That is, we need to find a family  $\{\tilde{\Psi}_k\}_{k \in \mathbb{Z}_+}$  of function in  $L_2^*([-\pi/2, \pi/2] \times [0, 2\pi])$  such that

$$\|f\|_*^2 = \sum_{j \in \mathbb{Z}_+} |\langle F, \tilde{\Psi}_j \rangle_*|^2,$$

where  $\|f\|_*$  and  $\langle f, g \rangle_*$  are the norm and the inner product associated with the weighted  $L_2$  space  $L_2([-\pi/2, \pi/2] \times [0, 2\pi])$ . The following result tells us how to do.

**Theorem 35** Suppose that  $\{\omega_j, j \in J\}$  is a tight frame for  $L_2(0, 2\pi)$  and  $\{\psi_k, k \in K\}$  is a tight frame for  $L_2(-\pi/2, \pi/2)$  with respect to a nonnegative weight function  $\cos(x)$ . Then  $\{\omega_j(\theta)\psi_k(\phi), j \in J, k \in K\}$  is a tight frame for  $L_2^*([-\pi/2, \pi/2] \times [0, 2\pi])$ .

*Proof* Since  $f \in L_2^*([-\pi/2, \pi/2] \times [0, 2\pi])$

$$\int_{-\pi/2}^{\pi/2} |f(\theta, \phi)|^2 \cos(\theta) d\theta$$

is essentially bounded by Fubini's theorem. For such a  $\phi \in [0, 2\pi]$  that

$$\int_{-\pi/2}^{\pi/2} |f(\theta, \phi)|^2 \cos(\theta) d\theta < +\infty,$$

we have

$$\begin{aligned} & \int_{-\pi/2}^{\pi/2} |f(\theta, \phi)|^2 \cos(\theta) d\theta \\ &= \sum_{k \in K} \left| \int_{-\pi/2}^{\pi/2} f(\theta, \phi) \psi_k(\theta) \cos(\theta) d\theta \right|^2 \end{aligned} \quad (112)$$

because that  $\{\psi_k, k \in K\}$  is a tight frame. Since

$$\begin{aligned} & \left| \int_{-\pi/2}^{\pi/2} f(\theta, \phi) \psi_k(\theta) \cos(\theta) d\theta \right|^2 \\ & \leq \int_{-\pi/2}^{\pi/2} |f(\theta, \phi)|^2 \cos(\theta) d\theta, \end{aligned}$$

it follows that  $\int_0^{2\pi} \left| \int_{-\pi/2}^{\pi/2} f(\theta, \phi) \psi_k(\theta) \cos(\theta) d\theta \right|^2 dy < \infty$ . For each  $k \in K$ , the function

$$\int_0^{2\pi} f(\theta, \phi) \psi_k(\theta) d\theta \in L_2(0, 2\pi).$$

The tight frame of  $\{\omega_j, j \in J\}$  implies that

$$\begin{aligned} & \int_0^{2\pi} \left| \int_{-\pi/2}^{\pi/2} f(\theta, \phi) \psi_k(\theta) \cos(\theta) d\theta \right|^2 d\phi \\ &= \sum_{j \in J} \left| \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} f(\theta, \phi) \psi_k(\theta) \cos(\theta) d\theta \omega_j(\phi) d\phi \right|^2. \end{aligned}$$

Therefore, integrating (112) from 0 to  $2\pi$  and using Lebesgue's dominant convergence theorem, we have

$$\begin{aligned} & \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} |f(\theta, \phi)|^2 \cos(\theta) d\theta d\phi \\ &= \sum_{k \in K} \int_0^{2\pi} \left| \int_{-\pi/2}^{\pi/2} f(\theta, \phi) \psi_k(\theta) \cos(\theta) d\theta \right|^2 d\phi \\ &= \sum_{k \in K} \sum_{j \in J} \left| \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} f(\theta, \phi) \psi_k(\theta) \cos(\theta) d\theta \omega_j(\phi) d\phi \right|^2 \end{aligned}$$

by (113). Therefore,  $\{\omega_j(\theta)\psi_k(\phi), j \in J, k \in K\}$  is a tight frame in  $L_2^*([-\pi/2, \pi/2] \times [0, 2\pi])$ .  $\square$

Let  $-\pi/2 = t_{1,0} < t_{1,1} < t_{1,2} < \dots < t_{1,m} = \pi/2$  be a knot sequence with  $t_{1,i} = -\pi/2 + ih$  with  $h = \pi/m$  and  $B_{1,i}$  be B-spline of order  $d > 1$  based on knots  $t_{1,i}, \dots, t_{1,i+d}$ , where  $t_{1,i+d} = t_{1,i+d-m}$  when  $i + d > m$ . Here, due to the equally-spaced knots  $t_{1,i}$ ,  $B_i(x)$  is just a shift of scaled version of uniform B-spline  $N_d$ . We uniformly refine the knots  $t_{1,i}, i = 0, \dots, m$  to have  $t_{2,i} = -\pi/2 + i\pi/(2m)$  and  $B_{2,i}$  be the B-spline of order  $d$  based

on knots  $t_{2,i}, \dots, t_{2,i+d}$ . Let  $V_1 = \text{span}\{B_{1,0}, \dots, B_{1,m-1}\}$  and  $V_2 = \text{span}\{B_{2,0}, \dots, B_{2,2m-1}\}$ . It is easy to see that  $V_1$  is refinable in the sense that  $V_1 \subset V_2$ . We repeat adding new knots and use the uniform B-splines on the new knots to form finer and finer spline spaces. Thus, we will have  $V_1 \subset V_2 \subset V_3 \subset \dots \subset L_2([-\pi/2, \pi/2])$ . It is known that  $\bigcup_{j=1}^{\infty} V_j$  is dense in  $L_2([-\pi/2, \pi/2])$  with respect to nonnegative weight function  $\cos(\theta)$ . Similar to the construction in Sect. “[Tight Wavelet Frames over Bounded Domain](#)”, we can find tight wavelet frames associated with these B-spline subspaces with respect to the weighted  $L_2$  inner product. For simplicity, let  $\{\psi_j, j \in J\}$  stands for the tight wavelet frame.

Similarly, we can find tight wavelet frame  $\{\omega_k, k \in K\}$  associated with trigonometric B-splines of even degree  $d > 0$ . See [53] for detail. Then  $\{\psi_j \omega_k, j \in J, k \in K\}$  form a tight wavelet frame by Theorem 35.

### Wavelets for Image Processing

In this section we explain how to use wavelets to find the edges of images, remove noises from images, and compress images. Mainly we use the wavelet decomposition and reconstruction as explained in Sect. “[Wavelet Decomposition and Reconstruction](#)”. That is, we first use a wavelet or a wavelet frame to decompose an image into several subimages. Then we treat these subimages by some methods. Finally we reconstruct the image back using treated subimages based on the same wavelet or wavelet frame. Certainly, different wavelets give us different reconstructed images. One of the purposes of the study is to find the best one from all wavelets constructed so far.

#### A Wavelet/Wavelet Frame Method for Edge Detection

We shall use tensor product of some orthonormal wavelets and the tight wavelet frame based on box spline  $B_{2211}$  to find the edges of images.

The wavelet/wavelet frame method for edge detection can be described as follows. We first use a wavelet or wavelet-frame to decompose an image into many levels of subimages which consist of a low-pass part and several high-pass parts of the image. Then we set the low-pass part to be zero and reconstruct the image back using zero low-pass part and the original high-pass parts. Such reconstructed image contains all the edges of the image. (See also [13].) For box spline wavelet frame, we only do one level of decomposition. For other standard wavelets (Haar, D4, D6, biorthogonal 9/7 wavelets), we do 1, 2, 3 levels of decomposition dependent on the images. For some images, e.g. the finger print image, we must do 3 levels of decomposition while for many other images, one or two

levels of decomposition are enough. We choose the best edge representation by visual inspection among three levels of decompositions to present here.

To present the edges clearly, we normalize the reconstructed image into the standard grey level between 0 to 255 and use a threshold to divide the pixel values into two major groups. That is, if a pixel value is bigger than the threshold, it is set to be 1. Otherwise, it is set to be zero.

In the following we present several sets of images for comparison. The top two figures are the original image and the edges based on the box spline  $B_{2211}$ . The two figures in the middle row are the edges computed by using the Haar and Daubechies D4 wavelets. The two figures in the last row are based on Daubechies D6 and wavelet 9/7 wavelet. From these figures, it is clear that box spline  $B_{2211}$  on four direction mesh does an excellent job to reveal the edges of images. (See more examples in [71].)

#### A Wavelet/Wavelet Frame Method for Image Denoising

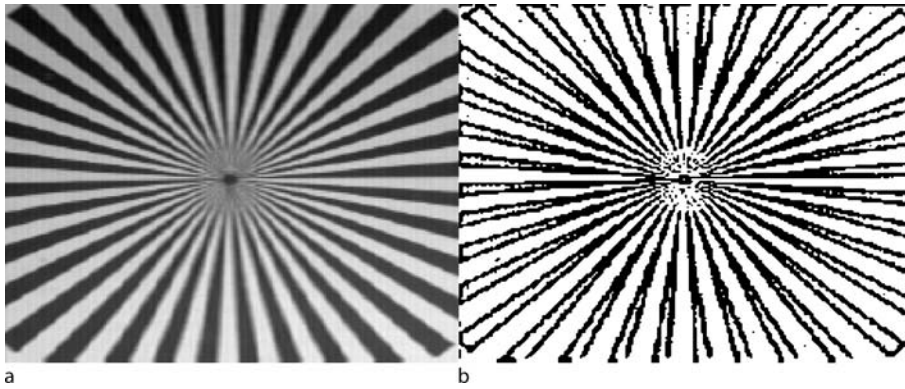
Here we shall use the tight wavelet frames based on box splines  $\phi_{111}, \phi_{221}, \phi_{222}, \phi_{1111}, \phi_{2211}$  to remove the noises from images. To compare the effectiveness of image denoising, we also use tensor products of the Haar wavelet, Daubechies D4 and D6 wavelets, and biorthogonal 9/7 wavelet to do the denoising. The main idea to use a wavelet or wavelet frame to denoise is we first decompose an image into one level of subimages which consist of a low-pass part and several high-pass parts of the image, then use the soft-thresholding [27,42] to treat each high-pass subimage by shrinking wavelet coefficients, and finally reconstruct the image by using the original low-pass part and shrunk high-pass parts. The reconstructed image is an denoised image.

The soft-thresholding method is to set each pixel value  $z$  of an image to be  $nz$  according to the following

$$nz = \begin{cases} 0, & \text{if } \text{abs}(z) \leq \epsilon \\ \text{sign}(z)(\text{abs}(z) - \epsilon), & \text{if } \text{abs}(z) > \epsilon \end{cases}$$

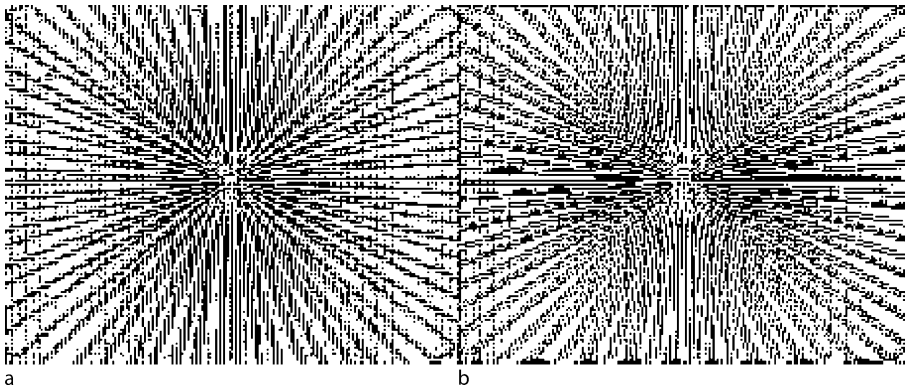
where  $\epsilon$  is a thresholding value. An heuristic reason for this wavelet denoising method is that the noise of the image will be in contents of the high-pass subimages after the wavelet decomposition. To remove the noise, we reduce the high frequency contents by  $\epsilon$ .

To measure the quality of denoised images, we use the peak signal to noise ratio (PSNR) which can be explained as follows. Let  $x_{ij}, 1 \leq i, j \leq 512$  be the pixel values of the original image with pixel values in the range  $[0, 255]$ . Let  $y_{ij}$  be the pixel value of the denoised image. Then we mea-



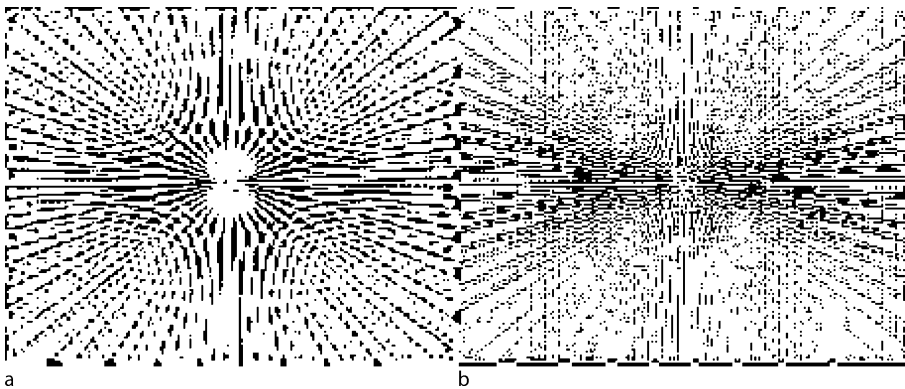
Popular Wavelet Families and Filters and Their Use, Figure 4

a Original image. b Edges by box spline  $B_{2211}$



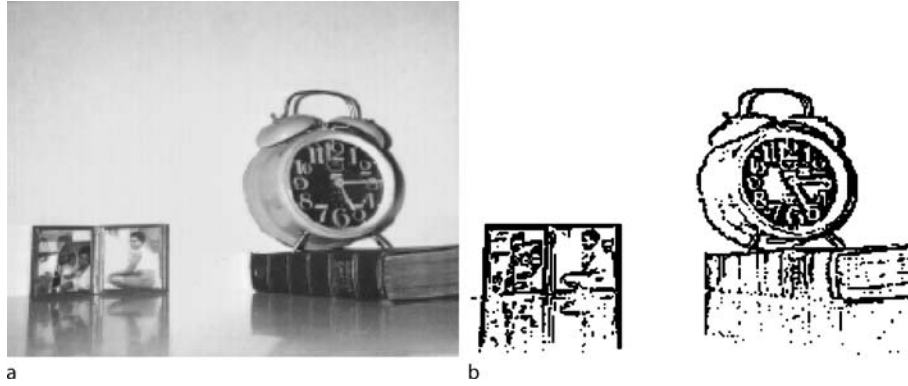
Popular Wavelet Families and Filters and Their Use, Figure 5

a Edges by Haar wavelet. b Edges by Daubechies D4 wavelet



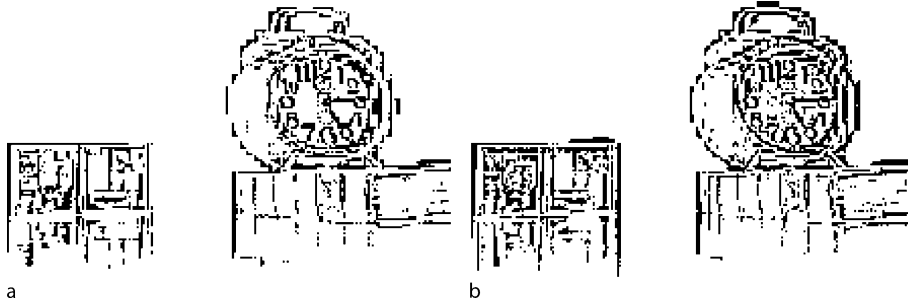
Popular Wavelet Families and Filters and Their Use, Figure 6

a Edges by D6 wavelet. b Edges by CDF 9/7 wavelet



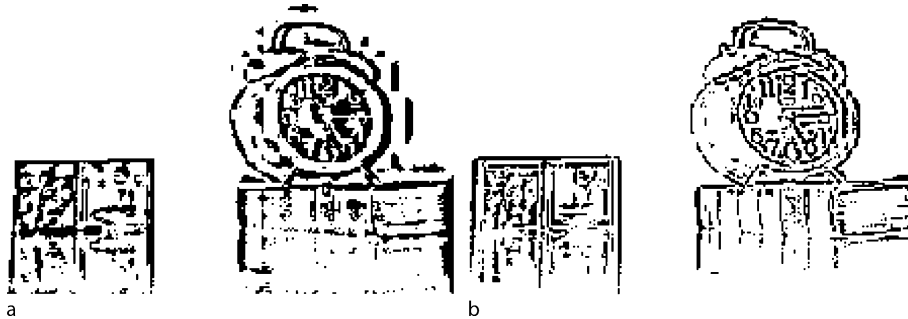
**Popular Wavelet Families and Filters and Their Use, Figure 7**

**a** Original image. **b** Edges by box spline  $B_{2211}$



**Popular Wavelet Families and Filters and Their Use, Figure 8**

**a** Edges by Haar wavelet. **b** Edges by D4 wavelet



**Popular Wavelet Families and Filters and Their Use, Figure 9**

**a** Edges by D6 wavelet. **b** Edges by CDF 9/7 wavelet

sure the PSNR of the denoised image by

$$\text{PSNR} = -10 \log_{10} \frac{\sum_{i,j=1}^{512} (y_{ij} - x_{ij})^2}{255^2 \times 512^2}.$$

Certainly, the thresholding value  $\epsilon$  is dependent on images and noises. We should look for an optimal  $\epsilon$  such that the PSNR is the best. It is known that the thresholding values are different for high-pass subimages at different levels of decomposition. It is also known that the most of the noise

will be in the high frequency contents at the first level of decomposition. one level of the wavelet decomposition.

To test the effectiveness of the wavelet/wavelet frame method, we start with an original image with pixel values  $x_{ij}$  and then add a Gaussian noise with zero mean and various variances  $\sigma$  to obtain a noisy image. That is, the pixel values of the noisy image is

$$z_{ij} = x_{ij} + \sigma \delta_{ij}$$



**Popular Wavelet Families and Filters and Their Use, Table 2**  
The PSNR comparison for standard image Bank

Images	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 25$
Noised	28.28	24.81	22.36	20.50
Haar wavelet	29.94	26.68	24.38	22.59
D4 wavelet	29.92	26.69	24.39	22.61
D6 wavelet	29.92	26.70	24.40	22.62
Biorth. 9/7	29.93	26.70	24.40	22.62
BSTF111	<b>31.59</b>	<b>28.99</b>	27.32	26.05
BSTF221	31.52	28.99	<b>27.40</b>	<b>26.24</b>
BSTF222	31.14	28.55	26.91	25.73
BSTF1111	31.34	28.78	27.20	26.04

with Gaussian noise  $\delta_{ij} \sim N(0, 1)$ , where  $\sigma$  will be chosen at levels 10, 15, 20 and 25. We choose a popular image Lena to test all the standard wavelets and box spline tight frames. In Table 2, we first list the PSNR numbers for noised image of Lena with different variances  $\sigma = 10, 15, 20, 25$  of noises. *BSTF111* stands for box spline tight frame using box spline  $\phi_{111}$ . Similar for other box spline tight frames. Then we present the noisy images and denoised images to show how well the box spline tight frames can perform.

### Wavelets for Image Compression

The scheme to use wavelets for image compression can be described as follows.

- (1) We use a wavelet to decomposing the gray-scale values of an images to a maximum number of levels.
- (2) Encoding the decomposed image using an embedded zero-tree encoder (cf. [81]) to a specified file size. For images of size  $512 \times 512$ , the file size of 262,159 bytes which is approximately one byte per pixel. The actual compressed file sizes are 32,793 bytes (8:1), 16,409 bytes (16:1), 8,217 bytes (32:1), and 4,121 bytes (64:1).
- (3) Decoding the compressed file.
- (4) Reconstructing the image using the wavelet transform and rounding the values to the nearest integer.
- (5) Calculating the peak signal to noise ratio (PSNR) as in the previous section.

The PSNR larger the better. That is, the compressed image can be recovered better. In general, a  $\text{PSNR} \geq 32$  is considered to be visually indistincted from the original and reconstructed images by normal people's eyes at a comfortable viewing distance away from the images. In the following we list a table of PSNR by using tensor prod-

uct of Haar, Daubechies D4, D6, and a Lai-Roach wavelet (cf. [59]). Test images are standard and given in Fig. 14.

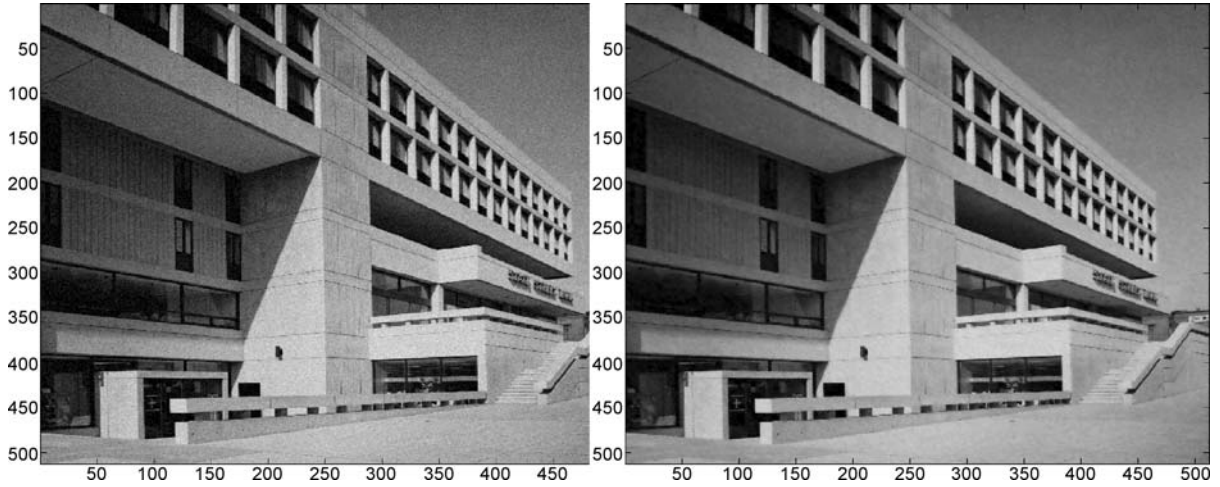
**Popular Wavelet Families and Filters and Their Use, Table 3**  
PSNR of five images by various wavelets

Image: Lena 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	36.2258	32.6462	29.5685	27.5420
D4	38.4440	34.9209	31.6733	28.8185
D6	38.7819	35.3234	32.0479	29.0727
9/7	39.6450	36.3379	32.9495	29.9049
LR6	38.8167	35.4208	32.1585	29.1588
Image: Barbara 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	30.4954	26.8119	24.6329	22.7409
D4	32.8675	28.6364	25.6853	23.3821
D6	33.4311	29.0735	25.9431	23.4715
9/7	34.7288	30.0708	26.4626	24.1701
LR6	33.6441	29.2804	26.1074	23.5789
Image: Boat 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	34.7720	30.7103	27.5762	25.4130
D4	35.6517	31.5910	28.5165	26.0746
D6	35.8593	31.8088	28.6402	26.1880
9/7	37.7118	33.1534	29.7557	27.1292
LR6	35.9301	31.9080	28.7187	26.2733
Image: Finger-print 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	32.4415	29.9961	28.5828	27.3396
D4	33.3224	30.8632	29.4448	27.9762
D6	33.8077	31.2097	29.8049	28.2562
9/7	34.7043	32.3327	30.3877	28.8153
LR6	33.9236	31.3387	29.9687	28.3928
Image: Crowd 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	33.3109	29.2032	26.1346	23.7222
D4	34.7214	30.6307	27.4101	24.8170
D6	35.1740	31.0444	27.7601	25.0348
9/7	37.0871	32.4109	28.8652	25.9603

In the above tables, D4 and D6 denote Daubechies wavelets of order 4 and 6, respectively. 9/7 stands for CDF 9/7 biorthogonal wavelet as in Sect. “[Biorthogonal Wavelets](#)”. As we can see that CDF 9/7 does much better for image compression for all levels. This is why that CDF 9/7 wavelets were implemented in JPEG 2000 standard.

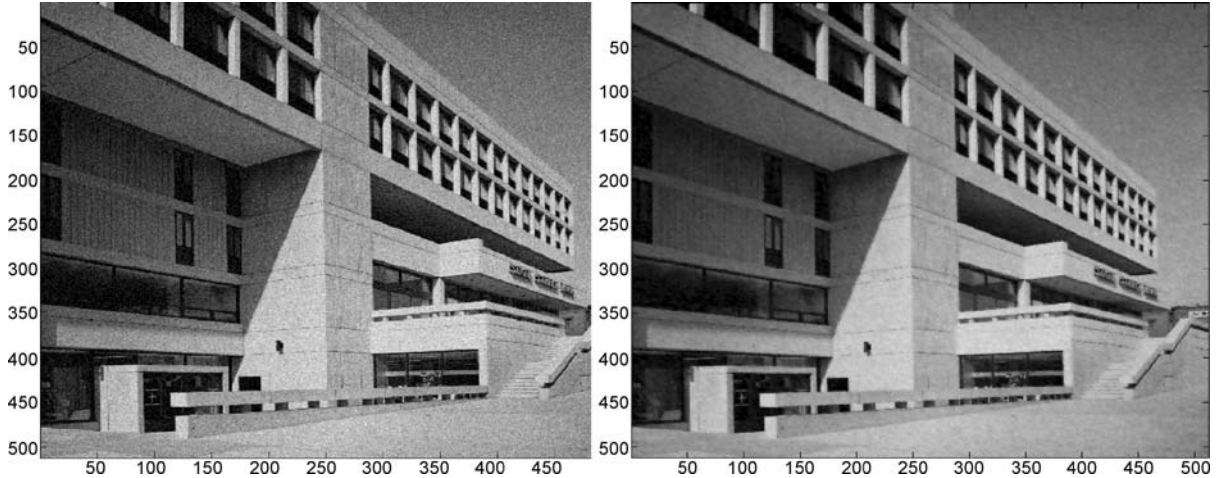
### Future Directions

We have discussed many constructive methods of various wavelets in the univariate and multivariate settings. There



Popular Wavelet Families and Filters and Their Use, Figure 10

The noisy image when  $\sigma = 10$  and a denoised image by using BSTF1111



Popular Wavelet Families and Filters and Their Use, Figure 11

The noisy image when  $\sigma = 15$  and a denoised image by using BSTF1111

are still many research problems remaining open after about twenty years of development of wavelets since the successful construction of compactly supported orthonormal wavelets with arbitrary regularity in 1988 [23]. We list some of important open problems below.

1. Construction of multivariate nonseparable compactly supported wavelets with high order smoothness. In the multivariate setting, how to find a mask polynomial  $P$  such that

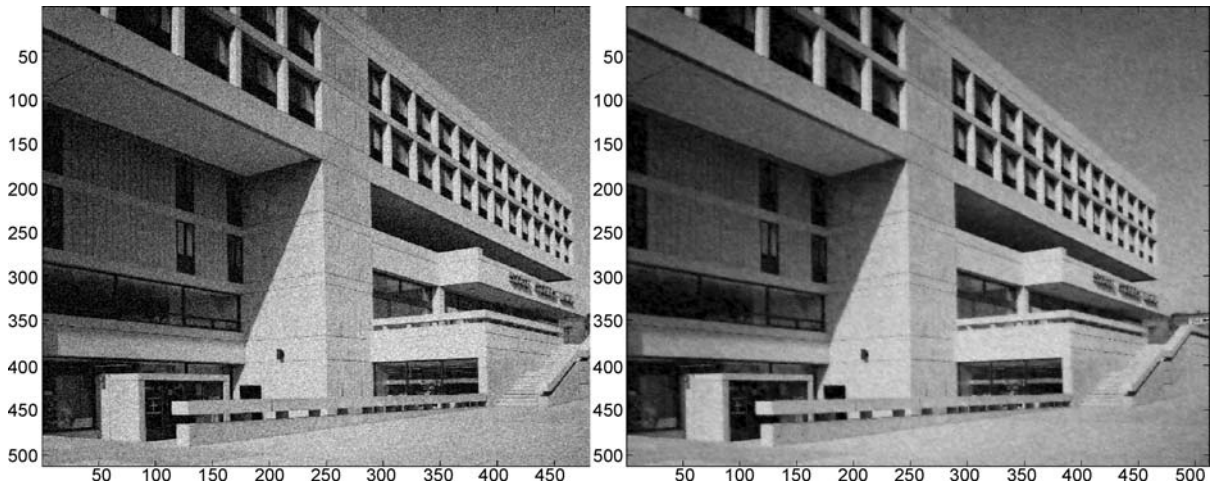
$$\sum_{k \in \{0,1\}^d} |P(\omega + k\pi)|^2 = 1$$

and  $\phi$  defined by its Fourier transform

$$\hat{\phi}(\omega) = \prod_{j=1}^{\infty} P(\omega/2^j) \quad (113)$$

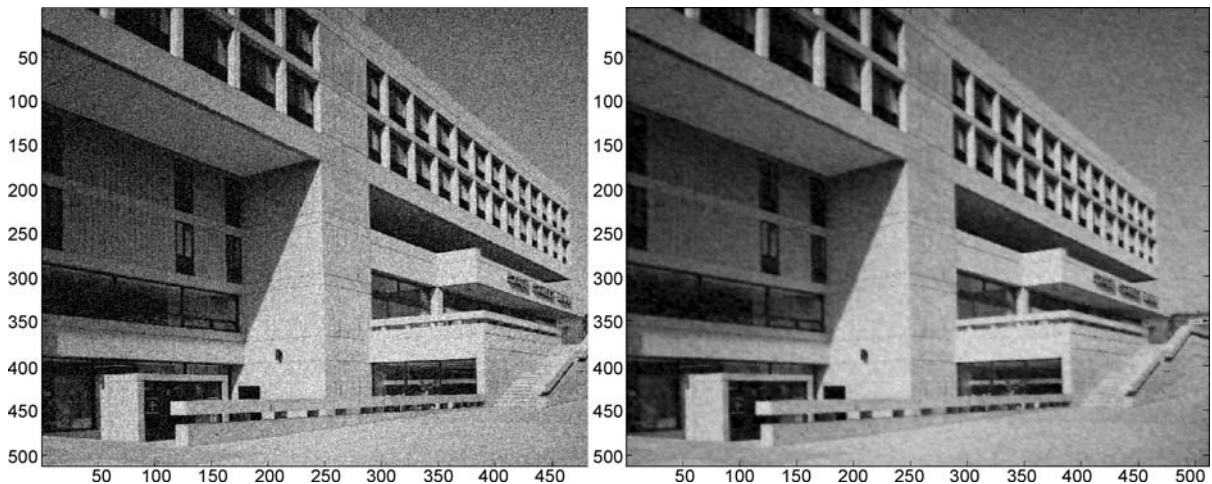
is compactly supported and has any desired smoothness. Belogay–Wang’s method in Sect. “[Multivariate Orthonormal Wavelets](#)” provides a way to do it in the bivariate setting. How can we do in the multivariate setting? Are there any other methods available?

2. In Sect. “[Multiwavelets and Balanced Multiwavelets](#)”, we describe the Goodman method for compactly supported orthonormal multi-wavelet using B-splines. One open problem is to generalize the Goodman



Popular Wavelet Families and Filters and Their Use, Figure 12

The noisy image when  $\sigma = 20$  and a denoised image by using BSTF2211



Popular Wavelet Families and Filters and Their Use, Figure 13

The noisy image when  $\sigma = 25$  and a denoised image by using BSTF2211

method to the multivariate setting. That is, can we use box splines to construct compactly supported orthonormal multi-wavelets?

3. Even in the univariate setting, when using B-splines to construct compactly supported orthonormal wavelets based on Goodman's method, it is interesting to know if we can use B-splines to construct symmetric compactly supported orthonormal multiwavelets.
4. In the univariate setting, consider using dilation factor  $q > 2$  to construct orthonormal wavelets. Suppose we have an orthonormal scaling function. That is, we have the mask polynomial  $P$ . It is known that we can use Lawton-Lee-Shen's method [62] to construct the associated wavelets. That is, we can find  $Q_k, k = 1, \dots, q-1$ . The open question is to find a formula for  $Q_k$ 's like the case when  $q = 2$ .
5. When constructing biorthogonal wavelets using box splines, the dual wavelets by the He-Lai method needs box splines of very high degree. Thus, the support of dual wavelets is very large. An open problem is to find multivariate biorthogonal wavelets of arbitrary smoothness with relatively small support. For example, for lower smoothness, say,  $C^r$  for  $r = 0, 1, 2$  how can we construct biorthogonal compactly supported wavelets of  $C^r$  with the support which is much smaller than the one by using the He-Lai method?
6. Suppose that there is a mask polynomial  $P$  satisfying (113). An open problem is if and how we can find





**Popular Wavelet Families and Filters and Their Use, Figure 14**  
**Images: a Barbara. b Finger-print. c Boat and d Crowd**

$Q_k, k = 1, \dots, 2^d - 1$  such that

$$\begin{bmatrix} P(\omega + n_1\pi) & P(\omega + n_2\pi) \\ Q_1(\omega + n_1\pi) & Q_1(\omega + n_2\pi) \\ \vdots & \ddots \\ Q_{2^d-1}(\omega + n_1\pi) & Q_{2^d-1}(\omega + n_2\pi) \\ \dots & P(\omega + n_{2^d}\pi) \\ \dots & P(\omega + n_{2^d}\pi) \\ \ddots & \vdots \\ \dots & Q_{2^d-1}(\omega + n_{2^d}\pi) \end{bmatrix}$$

is unitary, where  $\{n_k, k = 1, \dots, 2^n\} = \{0, 1\}^d \subset \mathbb{Z}_+^d$ .

7. Construction of multivariate compactly supported orthonormal wavelets, prewavelets, and tight wavelet frames in Sobolev spaces is a challenge for two decades. So far we still do not have a standard wavelet tool for numerical solution of partial differential equations. One initial step toward to this problem is [52]. Com-

pactly supported prewavelets under the norm in  $H_0^1(\Omega)$  with  $\Omega$  being a rectangular domain or triangular domain were constructed and tested to solve Poisson equation. More work is needed for a wavelet method for biharmonic equations and nonlinear equations.

8. As we have seen from the previous sections, tight wavelet frames are much easier to construct than orthonormal wavelets and they perform much better than orthonormal wavelets for image edge detection and denoising. However, one can not straightforwardly apply the wavelet compression schemes to do image compression based on tight wavelet frames. Mainly the number of wavelet frame generators is more than the number of wavelet functions and hence one needs more bit budget allocated to code the coefficients of wavelet frame representation of an image than to code the coefficients of orthonormal wavelet representation. This is one of major problems that wavelet researchers face today.

Regarding tight wavelet frames, one of problems is to increase the approximation power and order of vanishing moments of wavelet frames. This difficulty may be overcome by using the method of virtue components (cf. [55,56]). Construction of tight wavelet frames based on refinable function vectors will be reported in [51].

The wavelet research has already stimulated the interest from both pure and applied mathematicians. The problems proposed above require a deep knowledge of algebra and algebraic geometry (e.g., Problem 6) and an extensive knowledge of applied mathematics (e.g., Problem 8). It is not known if Problem 6 has a solution in the multivariate setting. The Problem 8 may need some knowledge of optimization and nonlinear sparse approximation. Thus, they are extremely difficult to solve. It may take time to further develop various linear and nonlinear approaches before we can see some hope to answer these questions.

## Bibliography

### Primary Literature

1. Ayache A (1999) Construction of non-separable dyadic compactly supported orthonormal wavelet bases for  $L^2(\mathbb{R}^2)$  of arbitrarily high regularity. *Rev Mat Iberoamericana* 15:37–58
2. Ayache A (2001) Some methods for constructing nonseparable, orthonormal, compactly supported wavelet bases. *Appl Comput Harmonic Anal* 10(1):99–111
3. Bastin F, Boigelot C (1998) Biorthogonal wavelets in  $H^m(\mathbb{R})$ . *J Fourier Anal Appl* 4:749–768
4. Bastin F, Laubin P (1997) Regular compactly supported wavelets in Sobolev spaces. *Duke Math J* 87:481–508
5. Belogay E, Wang Y (1999) Arbitrarily smooth orthogonal non-separable wavelets in  $\mathbb{R}^2$ . *SIAM J Math Anal* 30:678–697
6. de Boor C (1978) A practical guide to splines. Springer, New York
7. de Boor C, Hölig K, Riemenschneider S (1993) Box splines. Springer, New York
8. Chen G, Chui CK (1992) Signal processing and system theory. Springer, New York
9. Cho O, Lai MJ (2006) A class of compactly supported orthonormal B-Spline wavelets. In: Chen G, Lai MJ (eds) *Wavelets and Splines*. Nashboro Press, Brentwood, pp 123–151
10. Chirina M, Stöckler J (2007) Tight wavelet frames for irregular multiresolution analysis. *Appl Comput Harmonic Anal* (accepted for publication)
11. Chui CK (1988) *Multivariate splines*. SIAM Publications, Philadelphia
12. Chui CK (1992) *An introduction to wavelets*. Academic Press, San Diego
13. Chui CK (1997) *Wavelets: A Mathematical Tool for Signal Analysis*. SIAM Publication, Philadelphia
14. Chui CK, He W (2000) Compactly supported tight frames associated with refinable functions. *Appl Comp Harmonic Anal* 8:293–319
15. Chui CK, He W (2001) Construction of multivariate tight frames via Kronecker products. *Appl Comp Harmonic Anal* 11:305–312
16. Chui CK, Lian JA (1996) A study of orthonormal multi-wavelets. *Appl Numer Math* 20:273–298
17. Chui CK, Shi XL (1993) Bessel sequences and affine frames. *Appl Comp Harmonic Anal* 1:29–49
18. Chui CK, Wang JZ (1992) On compactly supported spline-wavelets and a duality principle. *Trans Amer Math Soc* 330:903–915
19. Chui CK, Stöckler J, Ward JD (1992) On compactly supported box-spline wavelets. *Approx Theory Appl* 8:77–100
20. Chui CK, He W, Stöckler J (2002) Compactly supported tight and sibling frames with maximum vanishing moments. *Appl Comp Harmonic Anal* 13:224–262
21. Cohen A, Daubechies I (1993) Nonseparable dimensional wavelet bases. *Revista Math Iberoamericana* 9:51–137
22. Cohen A, Daubechies I, Feauveau J-C (1992) Biorthogonal bases of compactly supported wavelets. *Commun Pure Appl Math* XLV:485–560
23. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Comm Pure Appl Math* 41:909–996
24. Daubechies I (1992) *Ten lectures on wavelets*. SIAM Publications, Philadelphia
25. Daubechies I, Han B, Ron A, Shen ZW (2003) *Framelets: MRA-based constructions of wavelet frames*. *Appl Comp Harmonic Anal* 14:1–46
26. Dierckx P (1986) An algorithm for fitting data over a circle using tensor product splines. *J Comp Appl Math* 15:161–173
27. Donoho D (1995) De-noising by soft-thresholding. *IEEE Trans Inform Theory* 41:613–627
28. Donovan GC, Geronimo JS, Hardin DP, Massopust PR (1996) Construction of orthonormal wavelets using fractal interpolation functions. *SIAM J Math Anal* 27:1158–1192
29. Donovan GC, Geronimo JS, Hardin DP (1996) Interwinning multiresolution analyses and the construction of piecewise polynomial wavelets. *SIAM J Math Anal* 27:1791–1815
30. Geronimo J, Lai MJ (2006) Factorization of multivariate positive Laurent polynomials. *J Approx Theory* 139:327–345
31. Goodman T (2003) A class of orthonormal refinable functions and wavelets. *Constr Approx* 19:525–540
32. Goodman T, Micchelli CA (1994) Orthonormal cardinal functions. In: Chui CK et al (eds) *Wavelets: Theory, algorithms and applications*. Academic, San Diego, pp 53–88
33. Han B (2007) Construction of wavelets and framelets by the projection method. *Int J Math Sci*, to appear
34. Hardin D, Hogan A, Sun Q (2004) The matrix-valued Riesz lemma and local orthonormal bases in shift-invariant spaces. *Adv Comput Math* 20:367–384
35. He W, Lai MJ (1997) Examples of bivariate non-separable continuous compactly supported orthonormal wavelets. *Proc SPIE* 3169:303–314. (It also appears in *IEEE Trans Image Proces* 9:949–953 (2000))
36. He W, Lai MJ (1999) Construction of bivariate compactly supported biorthogonal box spline wavelets with arbitrarily high regularities. *J Appl Comput Harmonic Anal* 6:53–74
37. He W, Lai MJ (2003) Construction of trivariate compactly supported biorthogonal box wavelets. *J Approx Theor* 120:1–19
38. Hong J, Lai MJ (2007) New constructions of orthonormal multiwavelets. Manuscript
39. Jia RQ, Micchelli CA (1992) Using the refinement equation for the construction of pre-wavelets. V. Extensibility of trigonometric polynomials. *Comput* 48:61–72



40. Jia RQ, Shen ZW (1994) Multiresolution and wavelets. *Proc Edinburgh Math Soc* 37:271–300
41. Jia R-Q, Wang J, Zhou D-X (2003) Compactly supported wavelet bases for Sobolev spaces. *Appl Comput Harmon Anal* 15(3):224–241
42. Johnstone I, Donoho D (1994) Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81:425–455
43. Karoui A (2003) A note on the construction of nonseparable wavelet bases and multiwavelet matrix filters of  $L_2(\mathbb{R}^n)$ , where  $n \geq 2$ . *Electron Res Announc Amer Math Soc* 9:32–39
44. Kotyczka U, Oswald P (1995) Piecewise linear prewavelets of small support. In: Chui CK, Schumaker LL (eds) *Approximation Theory VIII*, vol 2. World Scientific, Singapore, pp 235–242
45. Kovačević J, Vetterli M (1992) Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for  $(\mathbb{R})^n$ . *IEEE Trans Info Theory* 38:533–555
46. Kovačević J, Vetterli M (1995) Nonseparable two- and three-dimensional wavelets. *IEEE Trans Signal Proc* 43:1260–1273
47. Lai M-J (1992) Fortran subroutines for B-nets of box splines on three and four directional meshes. *Numerical Algo* 2:33–38
48. Lai M-J (2002) Methods for constructing nonseparable compactly supported orthonormal wavelets. In: Zhou DX (ed) *Wavelet analysis: Twenty year's development*. World Scientific, Singapore, pp 231–251
49. Lai M-J (2006) Construction of multivariate compactly supported orthonormal wavelets. *Adv Comput Math* 25:41–56
50. Lai M-J, Lian J-A (2007) A private communication on CDF9/7, May 2007
51. Lai M-J, Lian J-A (2007) Construction of tight multi-wavelet frames (under preparation)
52. Lai M-J, Liu HP (2007) Prewavelet solution to Poisson equation. Manuscript
53. Lai M-J, Lyche T (2007) Tight wavelets frames using trigonometric B-splines. Manuscript
54. Lai M-J, Nam K (2006) Tight wavelet frames over bounded domains. In: Chen G, Lai M-J (eds) *Wavelets and splines*. Nashboro Press, Athens, pp 313–326
55. Lai M-J, Petukhov A (2007) Method of virtual components for constructing redundant filter banks and wavelet frames. *Appl Comput Harmonic Anal* 22:304–318
56. Lai M-J, Petukhov A (2007) Method of virtual components in the multivariate setting. (submitted)
57. Lai M-J, Roach DW (1999) Nonseparable symmetric wavelets with short support. In: *Proceedings of SPIE Conference on Wavelet Applications in Signal and Image Processing VII*, vol 3813, pp 132–146
58. Lai M-J, Roach DW (2001) Construction of bivariate symmetric orthonormal wavelets with short support. In: Kopotun K, Lyche T, Neamtu M (eds) *Trends in approximation theory*. Vanderbilt University Press, Nashville, pp 213–223
59. Lai M-J, Roach D (2002) Parameterizations of univariate orthonormal wavelets with short support. In: Chui CK, Schumaker LL, Stoeckler J (eds) *Approximation theory X: Wavelets, splines, and applications*. Vanderbilt University Press, Nashville, pp 369–384
60. Lai M-J, Schumaker LL (2007) *Spline functions over triangulations*. Cambridge University Press, Cambridge
61. Lai M-J, Stoeckler J (2006) Construction of multivariate compactly supported tight wavelet frames. *Appl Comput Harmonic Anal* 21:324–348
62. Lawton W, Lee SL, Shen ZW (1996) An algorithm for matrix extension and wavelet construction. *Math Comp* 65:723–737
63. Lebrun J, Vetterli M (1998) Balanced multiwavelets: Theory and design. *IEEE Trans Signal Process* 46:1119–1125
64. Lyche T (1999) Trigonometric splines: A survey with new results. In: Peña J (ed) *Shape preserving representations in computer-aided geometric design*. Noa Science Publishers, New York
65. Lyche T, Schumaker LL (1994) L-spline wavelets. In: Chui C, Montefusco L, Puccio L (eds) *Wavelets: Theory, algorithms, and applications*. Academic Press, New York, pp 197–212
66. Lyche T, Schumaker LL (2000) A multiresolution tensor spline method for fitting functions on the sphere. *SIAM J Sci Comput* 22:724–746
67. Lyche T, Winther R (1979) A stable recurrence relation for trigonometric B-splines. *J Approx Theory* 3:266–279
68. Lyche T, Schumaker LL, Stanley S (1998) Quasi-interpolants based on trigonometric splines. *J Approx Theory* 95:280–309
69. Maass P (1997) Families of orthogonal two-dimensional wavelets. *SIAM J Math Anal* 27:1454–1481
70. Mallat S (1989) Multi-resolution approximations and wavelet orthonormal bases of  $L_2(\mathbb{R})$ . *Trans Amer Math Soc* 315:69–87
71. Nam K (2005) Box spline tight frames and their applications for image processing, Ph.D. Dissertation. University of Georgia, Athens
72. Petukhov A (2001) Explicit construction of framelets. *Appl Comp Harmonic Anal* 11:313–327
73. Petukhov A (2003) Symmetric framelets. *Constr Approx* 19:309–328
74. Riemenschneider S, Shen Z (1991) Box splines, cardinal series, and wavelets. In: Chui CK (ed) *Approximation theory and functional analysis*. Academic, Boston, pp 133–149
75. Riemenschneider S, Shen Z (1992) Wavelets and pre-wavelets in low dimensions. *J Approx Theory* 71:18–38
76. Ron A, Shen ZW (1997) Affine systems in  $L_2(\mathbb{R}^d)$ : The analysis of the analysis operator. *J Func Anal* 148:408–447
77. Ron A, Shen ZW (1997) Affine system in  $L_2(\mathbb{R}^d)$ , II. Dual systems. *J Fourier Anal Appl* 3:617–637
78. Ron A, Shen ZW (1998) Compactly supported tight affine spline frames in  $L_2(\mathbb{R}^d)$ . *Math Comp* 67:191–207
79. Ron A, Shen ZW (1998) Construction of compactly supported affine frames in  $L_2(\mathbb{R}^d)$ . In: Lau KS (ed) *Advances in Wavelets*. Springer, New York, pp 27–49
80. Rudin W (1963) The existence problem for positive definite functions. *Illinois J Math* 7:532–539
81. Said A, Pearlman WA (1996) A new fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans Circ Syst Video Technol* 6:243–250
82. Schoenberg IJ (1964) On trigonometric spline interpolation. *J Math Mech* 13:795–825
83. Schoenberg IJ (1973) Cardinal spline interpolation. SIAM Publication, Philadelphia
84. Schumaker LL (2007) *Spline functions, basic theory*. Cambridge University Press, Cambridge
85. Schumaker LL, Traas C (1991) Fitting scattered data on sphere like surface using tensor products of trigonometric and polynomial splines. *Num Math* 60:133–144
86. Simoncelli EP, Adelson EH (1990) Non-separable extensions of quadrature mirror filters to multiple dimensions. *Proc IEEE* 78:652–664

87. Stanhill D, Zeevi YY (1996) Two dimensional orthogonal wavelets with vanishing moments. *IEEE Trans Signal Proces* 46:2579–2590
88. Stanhill D, Zeevi YY (1998) Two dimensional orthogonal filter banks and wavelets with linear phase. *IEEE Trans Signal Proces* 46:183–190
89. Strang G, Fix G (1973) A Fourier analysis of the finite element variational method. In: Geymonat G (ed) *Constructive aspects of functional analysis*, C.I.M.E. II Ciclo 1971; 793–840
90. Vetterli M (1984) Multidimensional subband coding: some theory and algorithms. *Signal Proces* 6:97–112
91. Wickerhauser MV (1994) *Adapted wavelet analysis from theory to software*. AK Peters Ltd, Wellesley
92. Zhou J (2006) *Construction of orthonormal wavelets of dilation factor 3 with application in image compression*. Ph.D. dissertation, University of Georgia, Athens

### Books and Reviews

- Antoine JP, Murenzi R, Vanderghelynst P, Ali ST (2004) *Two-dimensional wavelets and their relatives*. Cambridge University Press, Cambridge
- Chen G, Lai M-J (2006) *Wavelets and Splines*. Athens, 2005. Nashboro Press, Brentwood
- Cohen A (2003) *Numerical analysis of wavelet methods*. North-Holland Publishing Co, Amsterdam
- Hernández E, Weiss G (1996) *A first course on wavelets, with a foreword by Yves Meyer*. CRC Press, Boca Raton
- Mallat S (1998) *Wavelet tour of signal processing*. Academic Press, San Diego
- MathWorks, Inc (2007) *MATLAB wavelet toolbox*
- Meyer Y, Coifman R (1997) *Wavelets, Calderon-Zygmund and multilinear operators*. Hermann, Paris, 1990 and Cambridge University Press, Cambridge
- Strang G, Nguyen T (1996) *Wavelets and filter banks*. Wellesley-Cambridge Press, Wellesley
- Vetterli M and Kovačević J (1995) *Wavelets and subband coding*. Prentice Hall, New Jersey
- Vidakovic B (1999) *Statistical modeling by wavelets*. A Wiley-Interscience Publication. Wiley, New York

## Positional Analysis and Blockmodeling

PATRICK DOREIAN

Department of Sociology, University of Pittsburgh,  
Pittsburgh, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Conventional Blockmodeling](#)

[Generalized Blockmodeling](#)

[Examples of Generalized Blockmodeling](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Social network** A simple social network is a set of social actors with one or more social relations defined over them. A network for one relation can be represented by a graph which is defined as an ordered triple  $G = (V, E, A)$ . In this representation,  $V$  denotes the set of vertices (or nodes) that represent the social actors,  $E$  the set of edges (undirected lines) representing reciprocated or symmetric ties and  $A$  the set of arcs (directed lines) representing unreciprocated or directed ties. The cardinality of  $V$ , usually denoted by  $n$ , gives the size of the network. The sets  $E$  and  $A$  are pairwise disjoint and their union is the set of all the ties for the social relation, denoted by  $R$ . The graph can also be denoted by  $G = (V, R)$  given that  $R = E \cup A$ . Such network data are also called one-mode data (because there is only one type of social unit over which social relations are defined). If there are  $r$  relations, these are denoted by  $R_1, R_2, \dots, R_r$  and the social network is denoted by  $G = (V, R_1, R_2, \dots, R_r)$ . For ease of exposition, a network with one relation is used below. The ties, where the arcs and edges are viewed as lines in the pictorial representation of a graph, can be binary or valued.

**Location in a network** The location of an actor ( $v_i$ ) is the set of ties that  $v_i$  has with all of the other actors in the network including the absence of ties.

**Partitioning networks** A partition of a social network is a simultaneous partitioning of the actors into *positions* and the social ties into *blocks*. For blockmodeling, this simultaneous partitioning is done solely in terms of the relational ties in  $R$ .

**Position** When the set of vertices ( $V$ ) is partitioned into a set of clusters  $C = (C_1, C_2, \dots, C_k)$  where  $C_i \cap C_j = \varnothing$ , the empty set, for distinct  $i$  and  $j$  and  $\cup_i C_i = V$ , then each  $C_i \in C$  (and the vertices it contains) forms a *position*. (See Fig. 2, below). The number of positions is denoted by  $k$ . This partition determines an equivalence relation and the units within each cluster are said to be equivalent to each other (and not equivalent to any the remaining vertices).

**Positional analysis** A positional analysis of a social network is based on the locations of all vertices (representing actors) and positions (occupied by vertices).

**Block**  $C$  also partitions the relation,  $R$ , into  $k^2$  blocks:  $R(C_i, C_j) = R \cap (C_i \times C_j)$  where  $\times$  denotes the Cartesian product. Each *block*,  $R(C_i, C_j)$ , consists of all of the arcs from vertices in  $C_i$  to vertices in  $C_j$ . The set of these arcs and edges is used to ‘summarize’ the relation between  $C_i$  and  $C_j$  by an arithmetic

calculation using the values of the arcs and edges in the block. A block is a relation between two clusters of vertices. (This is illustrated below in Sects. “[Introduction](#)”, “[Conventional Blockmodeling](#)” and “[Generalized Blockmodeling](#)” and diagrammed in Fig. 2). When  $i = j$  the blocks are called diagonal blocks and there are  $k$  of them, one for each position. The remaining  $k(k - 1)$  blocks are called nondiagonal blocks.

**Blockmodel** A *blockmodel* consists of structures obtained by: (i) identifying all vertices within a cluster (in the clustering  $C$ ) and representing each cluster as a vertex to construct  $k$  vertices for another graph and (ii) combining all of the ties in a block into a single tie between positions and constructing one tie for each block. If there are no ties in a block there is no tie between the two positions defining the block.

**Blockmodel image** The blockmodel is another network (graph) with many fewer vertices than  $G(k \ll n)$  and many fewer ties. When this blockmodel is represented as a simpler (and smaller) matrix or graph the result is labeled the *blockmodel image*. Even if there are no loops (ties from vertices to themselves) in  $G$  there can be loops in the blockmodel image depending on the ties present in diagonal blocks.

**Blockmodeling** Blockmodeling refers to the set of techniques used to discern network structure as blockmodels and representing them.

**Two-mode network data** A two-mode network  $N = (V_1, V_2, R, w)$  has one set of social units  $V_1 = (v_{11}, v_{12}, \dots, v_{1p})$  and a second set of units  $V_2 = (v_{21}, v_{22}, \dots, v_{2q})$  where  $V_1 \cap V_2$  is empty. The social relation  $R$  is a subset of  $V_1 \times V_2$  and is a relation *between* the two sets of vertices. The item,  $w$ , is a set of weights (viewed as a mapping of the relational ties of  $R$  to the real numbers).

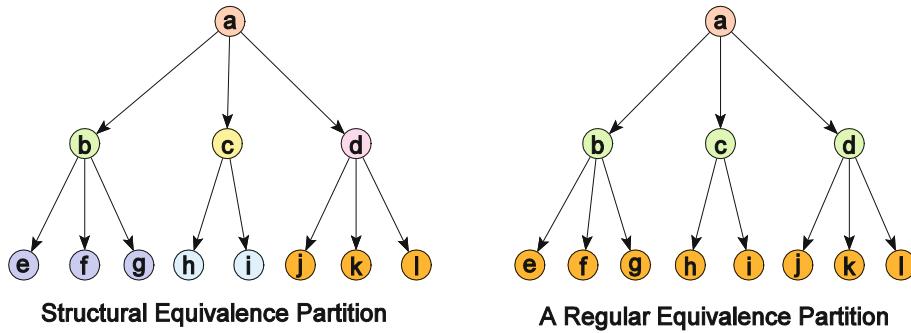
## Definition of the Subject

There are many types of social actors ranging from individuals to groups to organizations to institutions to societies. Many different social relations can be defined for all of these types of social actors. Depending on the substantive concerns of a researcher, many different and diverse social networks can be studied. All these networks can be represented by graphs as described above. It commonplace to talk of the structure of these networks: communication networks, transportation networks, organizational networks, trading networks all have structures. Two ideas are central to social network analysis. The first is that the structure of a social network, as a whole, is important for collective outcomes at the level of the network. The

second is that the location occupied *in* a network is important for outcomes at the actor level – regardless of the network that is studied. In order to study these collective and individual outcomes it is *essential* to know how a network is structured. Blockmodeling is an approach that is highly effective in identifying the overall structure of a network and describing it. Large and/or complex networks are mapped into simpler structures – called blockmodel images – that are viewed as structural summaries of the large and/or complex social networks. In one type of imagery, the structure depicted in a blockmodel image is the fundamental structure of the network and the observed network is an instantiation of the fundamental structure. Knowing the fundamental structure of a network is the primary goal. In turn, this permits the understanding of many network phenomena at both the collective and local levels.

## Introduction

The foundational paper for blockmodeling [19] introduced the concept of structural equivalence. Two actors (vertices) are structurally equivalent if they are connected to exactly the same other actors in the network. Their locations are identical. In this sense they are structurally identical and occupy exactly the same location in the network. Grouping together all of the vertices having the same network location defines a position occupied jointly by these (structurally equivalent) vertices. If all vertices can be grouped into sets of structurally equivalent clusters, the way is open to reduce a network to a simpler structure defined over these positions. The first widely accepted generalization of structural equivalence was regular equivalence as proposed in [21] where two vertices are regularly equivalent if they are equivalently connected to equivalent others. The difference between the two conceptions of equivalence is illustrated in Fig. 1 which shows a simple hierarchy twice. On the left, an exact structural equivalence partition is shown where the vertices have been color coded so that vertices in the same equivalence class have the same color. Note that in the lower two levels of the hierarchy not all of the vertices at a given level have the same color, a representation that appears to ignore a basic feature of this network. On the right, a regular equivalence partition is shown with three positions, one for each level. (As an illustration of the notion of regular equivalence, vertices  $j$ ,  $k$ , and  $l$  are connected to  $d$  in the same way as  $h$  and  $i$  are connected to  $c$  while  $c$  and  $d$  are connected in the same way to their subordinates.) There is an additional subtlety here because every network has a lattice of regular equivalence partitions [4]. Both of the



Positional Analysis and Blockmodeling, Figure 1

Two partitions of a hierarchy

partitions shown in Fig. 1 are located in the lattice of regular partitions of the network shown in the figure.

Of course, most empirical networks cannot be partitioned *exactly* in terms of structural or regular equivalence in a useful fashion. (Extremely fine grained partitions are possible for structural equivalence where  $k$  approaches  $n$  with the result that the blockmodel image is almost as large as the empirical network, an outcome that is not very useful. And, in terms of regular equivalence, there are the trivial partitions with 1 or  $n$  clusters for a network without isolates). So, in practice, partitions are sought that come as ‘close as is possible’ to the underlying conception of equivalence used to partition the network. In short, the notion of an exact equivalence partition is approximated when partitioning empirical networks to get a blockmodel. Conventional blockmodeling and generalized blockmodeling provide two approaches to establishing blockmodels empirically where the concept of an exact equivalence partition is approximated.

### Conventional Blockmodeling

Structural equivalence is defined as follows:  $x$  and  $y$  are structurally equivalent iff (i)  $xRy$  iff  $yRx$ ; (ii)  $xRx$  iff  $yRy$ ; (iii) For all  $z \in V \setminus \{x, y\}$ ,  $xRz$  iff  $yRz$  and (iv) For all  $z \in V \setminus \{x, y\}$ ,  $zRx$  iff  $zRy$ . In order to have a practical empirical procedure, conventional blockmodeling operationalizes an approximation to exact structural equivalence in terms of a metric. For example, if two vertices are structurally equivalent, the correlation of the vectors giving their locations is 1 because the vectors are identical. Similarly, the Euclidean distance between the two such vectors is 0. The correlation between “almost equivalent” locations will be close to 1 and the Euclidean distance between them will be close to 0. So “exactly equivalent” becomes “almost exactly equivalent” and clustering procedures can be applied to matrices containing

(dis)similarities that have been constructed from the relational data. This strategy has been called this ‘the indirect approach’ because the structural network data have been replaced by the (dis)similarity measures used for the clustering procedures [14]. One widely used algorithm, CONCOR [7], uses iterated correlations while another widely used algorithm, STRUCTURE [8], uses (corrected) Euclidean distances. Variants of both are implemented in UCINET [6].

An equivalence  $\approx$  on  $V$  is a regular equivalence on  $G = (V, R)$  iff for all  $x, y, z \in V$ ,  $x \approx y$  implies both (i)  $xRz$  implies there exists  $w \in V$  such that  $(yRw$  and  $w \approx z)$  and (ii)  $zRx$  implies there exists  $w \in V$  such that  $(wRy$  and  $w \approx z)$ . As for structural equivalence, regular equivalence does not apply exactly for most empirical networks. REGGE [20] has been used to relax exact regular equivalence in the form of an iterative algorithm to locate empirical partitions based on approximate regular equivalence. Thus far, no satisfactory metric for approximating exact regular equivalence has been established and, as a result, indirect methods for establishing partitions based on regular equivalence have been limited in their applicability. For a more general discussion of positions and equivalence, including automorphic equivalence, see [5].

### Generalized Blockmodeling

Generalized blockmodeling has been proposed as an alternative approach to conventional blockmodeling for establishing blockmodels empirically [14]. (See also [1]). While it shares the goal of discerning the structure of a network via homomorphisms of the network to the blockmodel image, it has many distinctive features: (i) it is a direct approach that works with the relational data; (ii) it generalizes the notion of equivalence in a way that permits a series of indefinite extensions; (iii) it uses an optimizational approach with an explicit criterion function; (iv) this cri-

terion function is used as measure of fit for established blockmodels; and (v) generalized blockmodeling can be used both inductively and deductively.

### Direct Approach

The direct approach is facilitated by translating the notion of equivalence into the set of block types that are consistent with the equivalence. Such block types are called *permitted* or *ideal* block types for the equivalence. Thus, structural equivalence is turned into the four block types that it implies: null block where every element is 0, complete blocks where every element is 1, diagonal blocks that have 0 only on the diagonal and 1 elsewhere, and diagonal blocks with only 1 on the diagonal and 0 elsewhere. (Of these block types, the fourth is empirically rare and is seldom used.) Under exact structural equivalence no other block types are possible. Regular equivalence is treated in a similar fashion: The only block types that regular equivalence permits are null blocks and 1-covered blocks [3]. A 1-covered block has at least one 1 in every row and every column. (Note that a complete block is a special case of a 1-covered block, consistent with structural equivalence being a special case of regular equivalence.) The implication of redefining equivalence in terms of a set of permitted block types means that an empirical procedure can be constructed by focusing on the extent to which the permitted block types are present in a network instead of transforming the structural data into (dis)similarities.

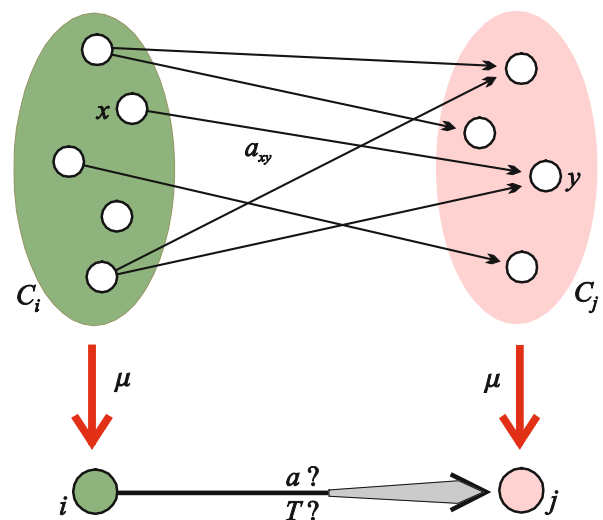
### Generalizing Equivalence Ideas

By focusing instead on permitted block types the way is cleared for defining new block types and new types of blockmodels. One way of doing this is to modify the definition of extant block types. The idea of 1-covered blocks can be relaxed from regular to row-regular blocks and column-regular blocks. The former have at least one 1 in every row while the latter has at least one 1 in each column. It follows that if a block is both row-regular and column-regular then it is regular. A row-dominant block is one where there is at least one row consisting entirely of 1s. If the clusters of the blockmodel partition are  $C_i$  and  $C_j$ , then at least one element in  $C_i$  has a tie directed all vertices in  $C_j$ . Column-dominant blocks can be defined in a similar fashion. (See pp. 211–215 in [14] for further examples of new block types).

New types of block types can be defined from scratch. An example of doing this stems from structural balance theory [17]. One generalization of Heider's theory [9] showed that if a signed network is structurally balanced

according to Heider, there will be a clear partition structure: the vertices will be partitioned into two clusters so that all of the positive ties will be within the clusters and all of the negative ties will be between the clusters. By modifying the definition of balance, this can be generalized so that there would be two or more clusters with the same partition structure [10]. This suggests a natural pair of block types and a new type of blockmodel [11]. The new block types are positive block (where the only elements are positive or null ties) and negative block (where the only possible ties are negative and null ties). According to the theory, the positive blocks will be on the diagonal and the negative blocks will be nondiagonal blocks. An algorithm for detecting partitions that are as close as possible to exact structural balance is suggested in [11]. Another example is the ranked clusters blockmodel where the diagonal blocks have vertices that are linked only by edges (or by null ties) and all of the arcs either go up the ranking or down the ranking [12]. Each of these blockmodels is an example of defining new block types and, with them, new types of blockmodels that can be fitted to empirical data.

The general scheme for generalized blockmodeling is shown in the diagram shown in Fig. 2. There, vertices in the cluster  $C_i$  are shown, generically, as  $x$  and the vertices of  $C_j$  are shown as  $y$ . The set of arcs from vertices in  $C_i$  to vertices in  $C_j$  is shown as  $a_{xy}$  in the figure. The cluster  $C_i$  of the original network is mapped under a mapping  $\mu$  to the position  $i$  (as a vertex in the blockmodel) and the cluster  $C_j$  is mapped to  $j$  (another position as



Positional Analysis and Blockmodeling, Figure 2  
A diagram of the generalized blockmodeling scheme



a vertex in the blockmodel) also under  $\mu$ . The set  $a_{xy}$  is summarized, via arithmetic rules, as  $a$  in the blockmodel and  $T$  is a set of predicates that are used to describe the block types of the blockmodel. Thus, for structural equivalence,  $T = \{\text{null}, \text{complete}\}$  and for regular equivalence  $T = \{\text{null}, \text{1-covered}\}$ . For structural balance,  $T = \{\text{positive}, \text{negative}\}$ .

### Criterion Functions

When permitted block types are defined, it is straightforward to define departures from the permitted (ideal) block types. For structural equivalence, every 1 in what is thought to be a null block is inconsistent with structural equivalence. Similarly, every 0 in what is thought to be a complete block is inconsistent with structural equivalence. The more inconsistencies there are in a block the less it conforms to the corresponding ideal block type. And the more inconsistencies there are, when the block inconsistencies are combined for the whole network, the worse the fitted empirical block model. Similarly, for regular equivalence, every row or column in a block lacking a 1 is not consistent with the block being 1-covered and every 1 in an otherwise null block is an inconsistency with regular equivalence. The more inconsistencies there are, when they are summed across the blocks, the worse the fitted blockmodel when regular equivalence has been used. Note that, in general, the count of inconsistencies for structural equivalence will differ from the count of inconsistencies for regular equivalence because the block types differ. Counting inconsistencies in blocks and combining them to give the total number of inconsistencies for a blockmodel is the core of the idea of a criterion function that can be used to establish blockmodels.

Let  $G = (V, R)$  be a network and let  $\Theta$  denote the set of all equivalence relations of a given type (for example structural equivalence, or regular equivalence, or balance theoretic, or ranked clusters). Every equivalence, say  $\sim$ , on  $V$  determines a partition  $C$  of  $V$  and vice versa. Let  $\Phi$  denote the set of all partitions corresponding to the relations from  $\Theta$ . In general, we need a criterion function, denoted  $P(C)$ , to satisfy two criteria: (i)  $P(C) \geq 0$  and (ii)  $P(C) = 0$  iff  $\sim \in \theta$ . With such a  $P(C)$  we can set up a clustering problem to minimize  $P(C)$  (given the equivalence specified for a blockmodel). If there are exact equivalences of a given type then  $P(C)$  is 0. However, if there are no such partitions (i.e.  $\Theta$  is empty) then an optimization approach gives the solution(s) that differ(s) the least from an ideal case – provided that the criterion function is compatible with and sensitive to the type of equivalence that is used.

The intuition behind using these criterion functions is that of a *pair of blockmodels* that are compared in terms of the criterion function. One is an ideal blockmodel with only the permitted block types for a given equivalence and the other is a corresponding empirical blockmodel (i.e. with the same partition of the vertices) for a clustering  $C$ . Let  $B(C_i, C_j)$  denote the set of all of the ideal blocks (in the blockmodel) and let  $p(C_i, C_j)$  denote the inconsistency between the empirical block defined by  $C_i$  and  $C_j$  and the corresponding nearest ideal block. The criterion function for the blockmodel is expressed as  $P(C) = \sum p(C_i, C_j)$  where the summation is over all  $k^2$  blocks. At the block level, the block inconsistency is  $p(C_i, C_j) = \min \delta(R(C_i, C_j), B)$  where the minimization is over all possible blocks for  $B(C_i, C_j)$  and  $\delta(R(C_i, C_j), B)$  measures the deviation (number of inconsistencies) between  $R(C_i, C_j)$  and the nearest ideal block  $B$ . For structural equivalence, the simplest inconsistency measure for a block whose nearest ideal block is null is the number of 1s in it. Put differently,  $\delta(R(C_i, C_j), B)$  is the number of 1s that are present where they ought not be. Similarly, the simplest inconsistency ( $\delta(R(C_i, C_j), B)$ ) for a block whose nearest ideal block is complete is the number of 0s in it. The total number of inconsistencies is obtained by summing the inconsistencies over all of the blocks in the blockmodel. This gives us  $P(C)$ . This can be generalized by using differential weights for the types of inconsistencies between the ideal and empirical blockmodels. For example, for structural equivalence partitions, it is possible to view 1s in what is specified as a null block as more serious than 0s in what is thought to be a complete block. If so, not all inconsistencies are viewed as being equally important and penalties can be imposed on some of them. Of course, this modifies the definition of, and values returned for, the criterion function,  $P(C)$ . The ‘best’ partitions  $C$  are not known ahead of time and have to be identified empirically. A simple relocation algorithm can be used determine one (or more) partitions that minimize the criterion function.

### Optimization and Measures of Fit

The relocation algorithm is defined and mobilized as follows. To locate a partition (as consistent as possible with the set of permitted block types defining the equivalence used) into  $k$  clusters, the network is partitioned randomly into  $k$  clusters. The value of  $P(C)$  is then computed. Of course, given that the starting partition is random, this value will be very large. The starting (and any) partition can be changed in two ways: (i) interchanging a pair of vertices between two different clusters or (ii) moving

a vertex from one cluster to another. These two types of change (transformations) determine the neighborhood of any clustering,  $C$ . The algorithm is expressed as follows:

Repeat:  
    if in the neighborhood of the current clustering,  $C$ ,  
        there exists a clustering,  $C'$ , such that  $P(C') < P(C)$   
    then move to the clustering  $P(C')$ .

This continues until a smaller value of  $P(C)$  cannot be found. The relocation algorithm has to be repeated many times (hundreds or thousands of times) for different random starting partitions for a given value of  $k$  in order to obtain solutions with the minimal values of  $P(C)$ .

A criterion function is defined *explicitly* for a particular type of blockmodel. As noted above, for regular equivalence there are two ideal block types. The value of  $\delta(R(C_i, C_j), B)$  is defined differently for each of these types. If  $B$  is a null block, then one specification of  $\delta(R(C_i, C_j), B)$  is the number of 1-covered rows and if  $B$  is a regular block,  $\delta(R(C_i, C_j), B)$  is the number of rows or columns that have only 0s. (Other specifications are possible.) For a structural balance theoretic partition, there are two block types – positive and negative blocks. For positive blocks (on the diagonal) every  $-1$  (for binary signed networks) contributes to  $\delta(R(C_i, C_j), B)$  and for negative blocks (off-diagonal) every  $+1$  (for binary signed networks) contributes to  $\delta(R(C_i, C_j), B)$  for that block. If  $P$  is the total number of ‘positive inconsistencies’ and  $N$  is the total number of ‘negative inconsistencies’ then one specification for the criterion function is  $P(C) = P + N$ . Alternatively, these two inconsistency types can be weighted differently as in  $P(C) = \alpha P + (1 - \alpha)N$  with  $0 \leq \alpha \leq 1$ . This formulation extends naturally to valued signed net-

works. A more extended discussion is contained in [13] and generalized blockmodeling is implemented in [3].

Deductive and Inductive Uses of Blockmodeling

Within conventional blockmodeling, using the indirect approach, blockmodeling is essentially inductive. A specific type of equivalence is specified, some (dis)similarity measure is specified (often implicitly) and a clustering algorithm is used to identify a clustering. For example, in using structural equivalence as the selected equivalence, corrected Euclidean distances can be computed and used in conjunction with the Ward criterion in a hierarchical clustering procedure. This usage results in a dendrogram or a cluster diagram and it is then necessary for a researcher to identify a clustering given the dendrogram. Neither the number of clusters is specified in advance nor is the location of the permitted types in a blockmodel specified. This is purely inductive. In another variant of using structural equivalence, CONCOR can be used to give successive splits of previously established clusters (starting with the whole network as a cluster) where the number of splits is specified in advance. This, too, is inductive and the researcher accepts (or not) the clusters that happen to be returned by the application of the algorithm. In contrast, generalized blockmodeling can be used in a pre-specified (and hence deductive use of blockmodeling) fashion. The different options are given in Table 1.

A simple pre-specified model is one where the block types and their location in the blockmodel are specified in advance but the actual clustering,  $C$ , is not known and is to be “discovered”. Two examples of pre-specified blockmodels for  $k = 3$  are given in Table 2 where both the types

Positional Analysis and Blockmodeling, Table 1  
Inductive and deductive uses of generalized blockmodeling

		Clustering	Blockmodel
Inductive		Unknown (in advance)	Unknown (in advance)
Deductive	Pre-specification	Unknown	Given
	Constraints	Given	Unknown
	Constrained Pre-specification	Given	Given

Positional Analysis and Blockmodeling, Table 2  
Two examples of pre-specified blockmodels

Model 1			Model 2		
Complete	Complete	Null	Row Regular	Row Dominant	Complete
Complete	Regular	Null	Null	Complete	Null
Regular	Null	Null	Null	Regular	Symmetric

of blocks and their place in the blockmodel are specified. These specifications state that the type of blockmodel is known or conjectured but not its exact realization.

It is permissible to formulate alternative blockmodels for the same network as alternative hypothetical models. However, the two criterion functions cannot be compared directly (in magnitude) because different block types define different inconsistencies. It is possible – and interesting – for different blockmodels to fit the same network.

In general, the criterion function for structural equivalence is the most stringent with every 0 and 1 in the ‘wrong’ places (blocks of the ‘opposite’ type) contributing while the number of inconsistencies under regular equivalence is likely to be much smaller. So comparing the criterion functions for different blockmodels based on structural and regular equivalence for the same data is inadmissible. In short, the metrics differ and the size of the criterion function matters only within a *given* blockmodel type for a specific network data set – where the aim is to find the closest empirical blockmodel(s) consistent with the selected type of blockmodel for the network studied. What matters is the appropriate type of blockmodel given the problem studied. The use of criterion functions is to identify the best fitting blockmodels of a given type and they should never be compared across different types of blockmodels.

An alternative option is to specify the clustering and leave open the exact nature of the blockmodel. The constraint takes the form of specifying which vertices go in which cluster and, given that, seeing the blockmodel type that results (in terms of the permitted block types). In the first example used below the presence of two types of organizations in the network suggested a partition of the organizations into the two types. Under this strategy, it is possible also to identify only partially the membership of the clusters by specifying that some vertices must be clustered together and other pairs of vertices are never clustered together. The third approach listed in Table 1 is the constrained pre-specification where both the blockmodel and the clustering are specified. (In the organizational example just mentioned, the partition into two types of organizations could be coupled to a specification of one type of a core-periphery blockmodel.)

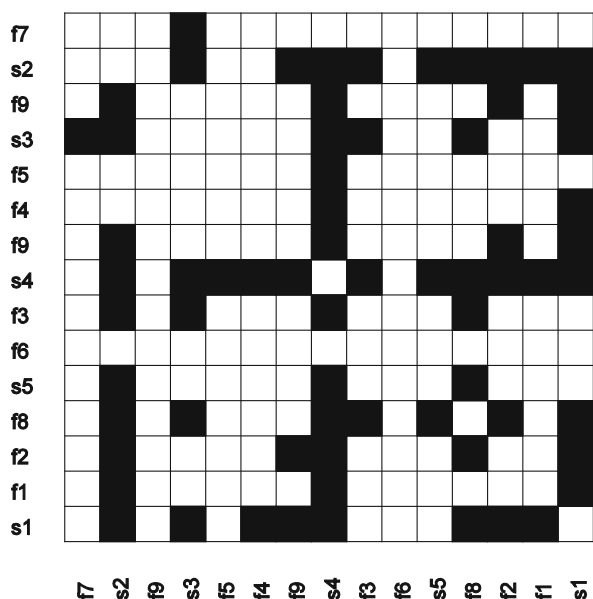
Engaging in pre-specifying blockmodels requires social knowledge about the specific network or the type of network that it represents. And it may be that a lot of such knowledge is required to make blockmodel specifications with confidence. However, often a network analyst knows more about a network than what is implied by a completely inductive approach. As such, a pre-specified blockmodel is one that can be *tested* rather than one

that is discovered. All of the inductive, partially deductive and fully deductive uses of blockmodeling are useful and the decision over which approach to take depends on the knowledge that is available about a network. For example, a type of blockmodel that is discovered for a network of a given type (e. g. corporate networks in the United States) becomes a candidate for a blockmodel to be tested in another network of the same broad type in another country.

## Examples of Generalized Blockmodeling

### Example 1. An Interorganizational Network

The first example features a small network with 15 organizations and is taken from [15]. The studied organizations are involved in a specific collaborative enterprise and the relation shown in Fig. 3 is the presence (or not) of reciprocated ties prior to the start of the collaborative enterprise. Even though this is a small network, the structure is not immediately apparent from the diagram. Prior to the analysis, the basic intuition was that these organizations form some kind of a core-periphery structure. Of course, there are many types of core-periphery structures so something more has to be specified. Those organizations thought to be in the core were expected to be tied to each other and to most of the other organizations in the network. These organizations occupy the core as Position 1. Another set of organizations was each thought to be tied to most of the core organizations but not to each other. These organiza-



Positional Analysis and Blockmodeling, Figure 3

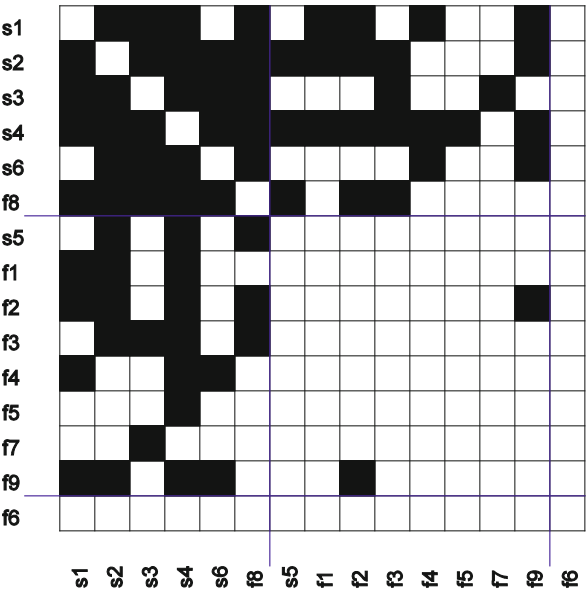
An interorganizational network matrix in an arbitrary order

Positional Analysis and Blockmodeling, Table 3  
A pre-specified blockmodel

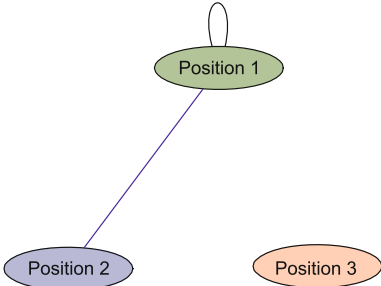
	Position 1	Position 2	Position 3
Position 1	Complete	Regular	Null
Position 2	Regular	Null	Null
Position 3	Null	Null	Null

tions occupy Position 2 and can be called the semi-periphery. Finally, there was the suspicion that one organization largely unknown prior to the collaborative effort. This organization (or possibly more than one organization) occupies Position 3. Reflecting these arguments, the specification used in [15] is shown in Table 3.

This pre-specified model was fitted with 3 clusters and the rows of the matrix were permuted so that members of the same position (cluster) are grouped together. The matrix array in Fig. 4 resulted. The blue lines (that extend beyond the matrix on the left and at the bottom) separate the three positions. While the model fits the data well, there are some inconsistencies: among the core organizations in Position 1, s6 and s1 do not have a reciprocated tie (and contribute 2 inconsistencies to the criterion function as 0s in a specified complete block) and, among the organizations in the semi-periphery (Position 2), f9 and f2 do share a tie when the expectation expressed in the model was there would be no ties there. This brings the number of inconsistencies to 4. There is a lone organization in Position 3 that has no reciprocated ties with the other



Positional Analysis and Blockmodeling, Figure 4  
Blockmodel of the interorganizational network with 3 positions



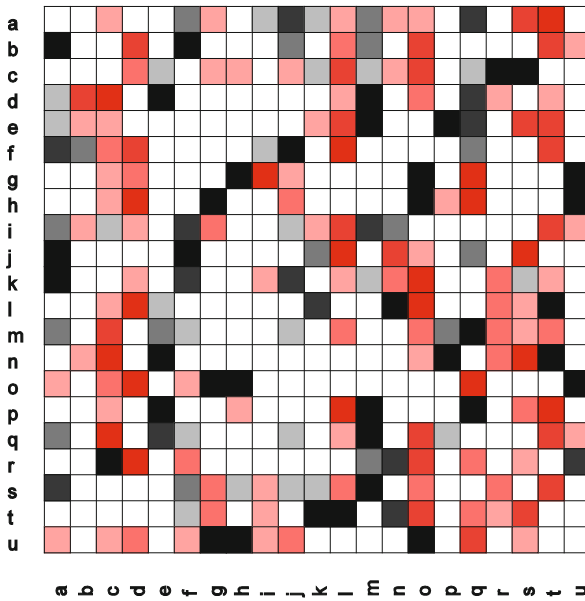
Positional Analysis and Blockmodeling, Figure 5  
Blockmodeling image for Fig. 4

organizations: the blocks defined in terms of this position contribute no inconsistencies. So the optimized value of the criterion function is 4 and this partition is unique. Figure 5 shows the blockmodel image for this network gotten by using the pre-specified model shown in Table 3. The loop for Position 1 represents a complete block and the edge between Position 1 and Position 2 represents a regular link.

This example raises another interesting issue. As noted above in Sect. “Conventional Blockmodeling”, the organizations could be grouped into two clusters based on what they do. One was made up of ‘practicing’ organizations (with regard to health service delivery) and the other was made up of organizations ‘supporting’ the collaborative effort. The former organizations are labeled f1 through f9 and the second set s1 through s6. An alternative blockmodel was an obvious candidate and is consistent with the third deductive use of blockmodeling listed in Table 1. It has the same specification of block types with three positions: Position 1 made up of the support organizations; Position 2 made up of all of the practicing organizations. (Position 3 could be specified for f6 as an isolate.) Much as this blockmodel made sense a priori, it has a much worse fit. Looking at Fig. 4, it is clear that if s5 and f6 are interchanged between Positions 1 and 2, the count of inconsistencies would jump by 12 to 16. The structure of the second specification is right but the composition of the positions was inferior – which shows, for this network, that the constrained pre-specification did not fit as well as a simpler pre-specification with only the blockmodel stated ahead of the model fitting. It suggests also that *testing* specifications is important and that the assumed social knowledge for specified blockmodel need not be correct.

Example 2. Structural Balance and Signed Networks

The data for this example are taken from [18] for a group of 21 women living in a college residence. Measurement for signed social relations was done by using four rela-



Positional Analysis and Blockmodeling, Figure 6  
Original signed and valued matrix

tions that tapped an underlying dimension of affect. Each woman was asked to provide signed data about the other women in the residence in the form of who they would like to do activities with and those with whom they did not want to do these things. The items used were: having them as a room mate; going on a double date with them; taking them home to visit the family for a weekend and being a friend after college. These relations have been combined into counts of activities to be shared and counts of activities not to be shared are shown in Fig. 6.

Black squares indicate the presence of positive ties for all four relations with red squares showing where all four ties were negative. Shades of gray show lesser counts of positive ties (+3, +2 and +1) and shades of pink show lesser counts of negative ties (−3, −2 and −1). White squares indicate no preference either way. (There were no cases where, for a pair or women, there was a positive tie on one relation and a negative tie on another relation). If structural balance holds exactly then a signed relation will partition the actors into clusters (also called plus-sets) where all of the ties within a cluster are positive and all ties between the clusters are negative. In practice, structural balance does not hold exactly and there are some negative ties within plus-sets and some positive ties between them. These are all inconsistencies with structural balance and the total count of them gives the value of a criterion function,  $P(C)$ , and this depends on the number of clus-

Positional Analysis and Blockmodeling, Table 4

Values of  $P(C)$  and the number of solutions for values of  $k$

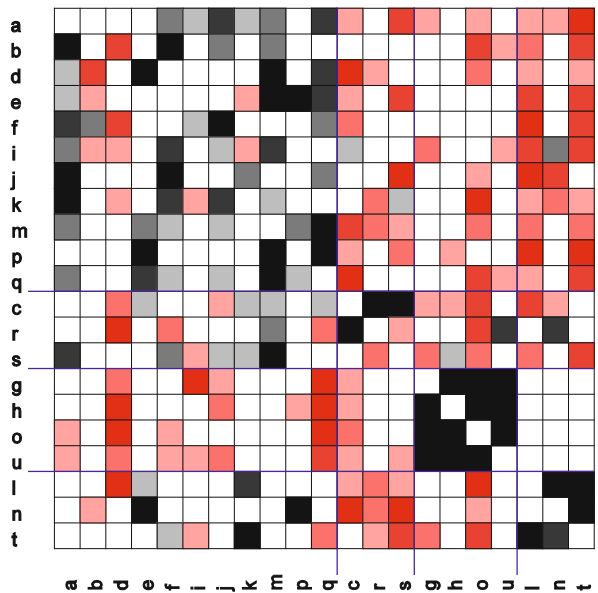
$k$	2	3	4	5	6
$P(C)$	48.5	37	32	32.5	33
Solutions	1	1	1	2	1

ters ( $k$ ) used. Table 4 summarizes the outcomes leading to the choice of the partition shown in Fig. 6. (The criterion function,  $P(C)$ , is larger for higher values of  $k$ ).

The best fitting (and unique) partition is with 4 clusters. The pre-specified blockmodel for this case is:

Positive (+, 0)	Negative (−, 0)	Negative (−, 0)	Negative (−, 0)
Negative (−, 0)	Positive (+, 0)	Negative (−, 0)	Negative (−, 0)
Negative (−, 0)	Negative (−, 0)	Positive (+, 0)	Negative (−, 0)
Negative (−, 0)	Negative (−, 0)	Negative (−, 0)	Positive (+, 0)

To locate the best fitting partition(s), the relocation algorithm described above was used with  $P(C) = \alpha P + (1 - \alpha)N$  with  $\alpha = 0.5$ . The unique best fitting partition is shown in Fig. 7 where the blue lines mark the boundaries between the four plus-sets. Any red or pink square in the diagonal blocks shows the presence of negative ties, in what is specified as positive blocks, as inconsistencies contributing to  $P(C)$ . Similarly, black and gray squares in the off-diagonal blocks show positive ties that are present in what are specified as negative blocks and they also con-



Positional Analysis and Blockmodeling, Figure 7  
Permuted signed and valued matrix with 4 plus-sets



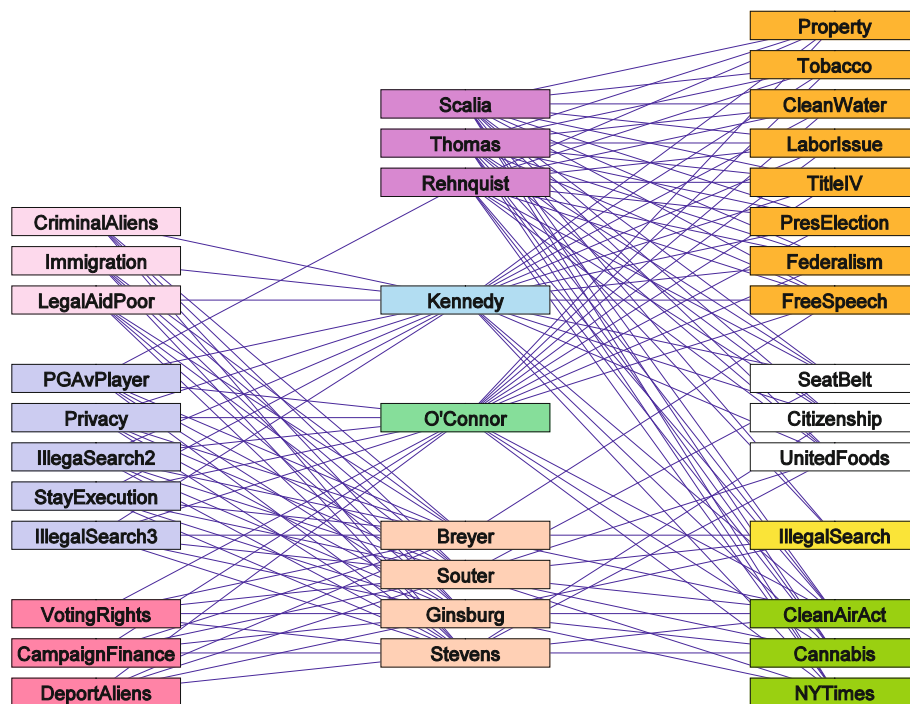
tribute to  $P(C)$ . The plus-set  $\{g, h, o, u\}$  is particularly cohesive with all their positive ties being distributed among themselves and all their negative ties being sent to women in other plus-sets. Furthermore, the women in this plus-set receive no positive ties from women in other plus-sets. The plus-set  $\{l, n, t\}$  has only positive ties within their plus-set but two positive ties are sent to women in other plus-sets. All of the negative ties are sent outside the plus-set. The other three person plus-set is less consistent with the ideal structural balance partition having two negative ties within it and positive ties to women in other plus-sets. The blocks associated with the large plus-set have both kinds of inconsistencies. In summary, the generalized blockmodeling discerned a structural pattern as a partition of a valued signed relation that is readily interpretable.

### Example 3. The Supreme Court as Two-Mode Data

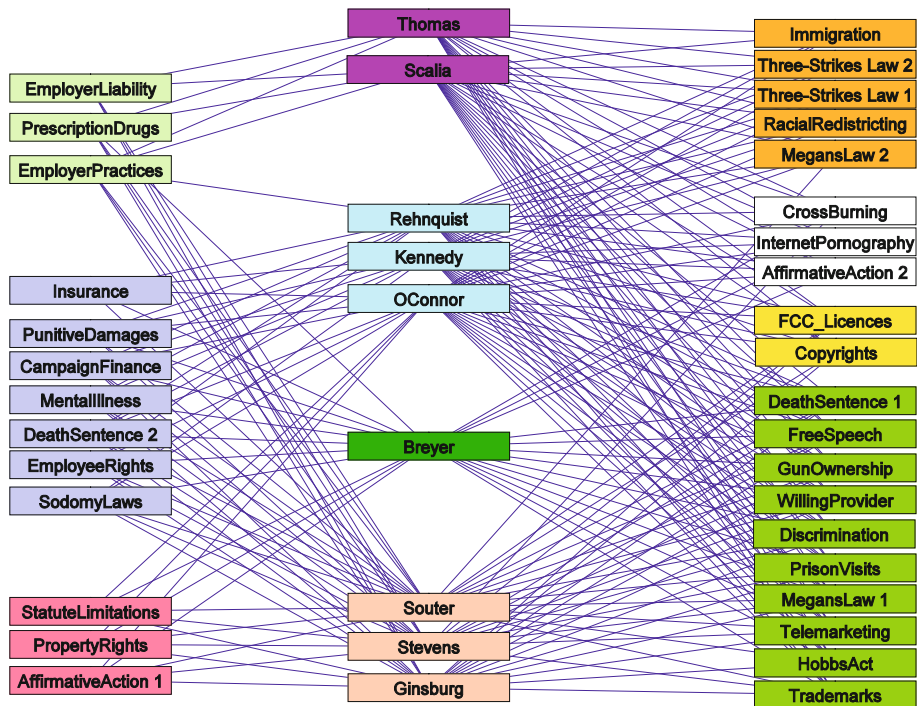
Two-mode data involves two sets of social objects. Examples of such data include companies employing people, people attending events, organizations joining alliances and Supreme Court Justices voting on cases they hear. While the row objects and column objects belong to different sets, it is still possible to apply blockmodeling tools to two-mode data. Of course, the rows and columns will

be partitioned differently. For the example of the voting of Supreme Court Justices there are two possible weights: 1 for voting in favor of a decision and 0 for not so voting. (An alternative weighting scheme has three values where 1 is voting for a decision, 0 is not voting at all (for justices recusing themselves from a case) and  $-1$  for voting in dissent). The simpler weight scheme is used in this example.

Instead of seeking a clustering  $C$  as used for one-mode data, a two-clustering is sought where  $C = (C_1, C_2)$  with  $C_1$  a partition of  $V_1$  and  $C_2$  a partition of  $V_2$ . As for one-mode data, block types, together with their associated inconsistency counts, can be defined and an optimization problem formulated as  $(\Phi, P(C), \text{minimize})$ . The solution set is made up of two-clusterings  $C^* = (C_1^*, C_2^*)$  for which  $P(C^*)$  is minimized over all  $C \in \Phi$ , the set of feasible two-clusterings. The value of the criterion function,  $P(C^*) = P(C_1, C_2)$  is obtained by summing all of the inconsistencies for each block of the two-mode blockmodel. Structural equivalence blocks (null and complete) was used in [14] to examine the voting of the Supreme Court for the 2000–1 term for the 26 most important decisions identified in [16](a). The justices were partitioned into 4 clusters and the cases into 7 clusters with 13 inconsistencies. Their optimal solution partition is shown in Fig. 8 where the four clusters of justices are in the center

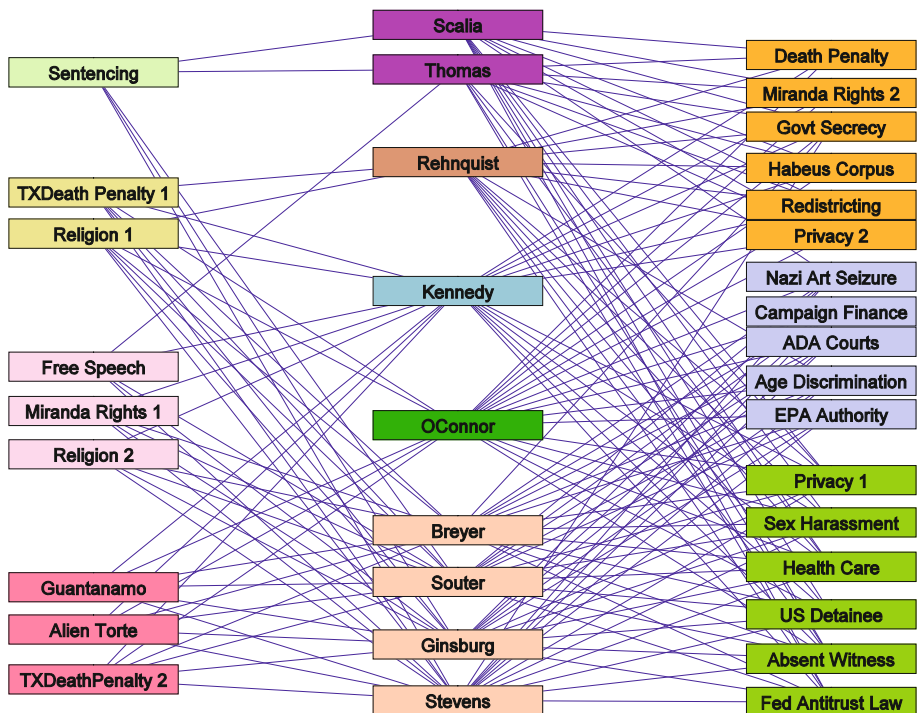


Positional Analysis and Blockmodeling, Figure 8  
A partition of the Supreme Court voting for 2000–2001



Positional Analysis and Blockmodeling, Figure 9

A partition of the Supreme Court voting for 2002–2003



Positional Analysis and Blockmodeling, Figure 10

A partition of the Supreme Court voting for 2003–2004

column of the figure and the clusters of cases are shown on the two sides. Both sets of clusters are color coded according to cluster membership. The blue lines show the votes of the justices for their decisions. The partition of the justices shows the familiar ideological cleavage of the Supreme Court with the wing of ‘conservative’ justices at the top and the ‘liberal’ wing at the bottom.

Figures 9 and 10 show the corresponding results for the next two years of the Supreme Court’s decisions based on the important cases identified in [16](b, c). There are subtle differences between the three two-mode blockmodel partitions within the same broad description of the voting patterns of the Supreme Court across the years. The methodological point here is that generalized blockmodeling can be applied to two-mode data and that the resulting partitions are both clear and interpretable.

The benefits from using a generalized blockmodeling approach include the use of precise criterion functions and an optimization approach for identifying blockmodels empirically, having a well defined measure of fit for a blockmodel, being able to define new types of blocks and new blockmodel types, being able to use blockmodeling in a deductive fashion (when our knowledge merits doing so), and being able to differentially weight the different types of inconsistencies in blocks and doing so within a very flexible, coherent and broad framework for blockmodeling. Further, in cases where the direct and indirect approaches have been compared, for both structural and regular equivalence, the generalized approach has never been outperformed by the indirect approach and most often it returns blockmodels that fit better for the same equivalence type. However, even though these benefits are clear, it does not follow that the generalized approach to blockmodeling totally dominates the conventional approach nor that the major problems have been solved.

### Future Directions

One drawback to generalized blockmodeling is that the combinatorial computational burden is considerable even when a local optimization method is used. This constrains the size of networks that can be modeled, a major disadvantage. Much larger networks can be analyzed using the indirect approach and, even though attention is confined to structural and regular equivalence with it, the network size restriction for generalized blockmodeling is problematic. This issue becomes even more acute when three-mode networks are considered. For these problems, some combination of indirect methods, direct methods and graph theoretical methods will be needed. So, one

open problem set stems the need to create more efficient algorithms together with the formulation of better heuristics for partitioning networks in general.

Positional analysis gives priority to network locations and network positions. Indeed, this is its hallmark. Because the whole network matters, both approaches to blockmodeling assume that the boundary of the network has been located correctly. In many cases, the assumption is appropriate but, as the ambition of network analysts expands to consider networks of (much) larger size, this assumption becomes questionable. At the moment, we simply do not know the vulnerability (or instability) of blockmodeling solutions when the boundary has been drawn inaccurately. (Leaving out crucial network vertices is far more problematic than including vertices of little structural relevance). The need to learn the vulnerabilities of blockmodeling to this type of boundary problem is acute and creates a second open problem set. One data analytic response to this problem is to represent the wider network environment of the network being studied in some fashion so as to reduce the vulnerability that comes from simply ignoring the wider network environment.

Equally acute is the problem of missing network data when the boundary has been correctly identified. That is, the data may contain all of the actors but the information from and about them is incomplete, inaccurate or missing. This sets up the third set of open problems: dealing with inadequate data. The missing data part of this problem is easy to sweep under the rug by assuming that the missing relational data is the same as null relational data. In reality, the 0 in a network data set can be a null relation or that no data for the tie is available and it is folly to treat them as the same. There are three broad approaches to this problem that show promise. One is to experiment with complete extant data by systematically deleting or otherwise corrupting the data and reanalyzing the corrupted data with the same blockmodeling tools to learn more about which types of corrupted data matter the most and how much missing or corrupted data will distort the results of blockmodeling the network. A second approach is to use simulation methods to generate data where the sources of missing or incomplete data is built in the generating process and can be studied. The disadvantage to both of these approaches is that the derived knowledge may be restricted to the situations featured in the experiments and simulations. The third approach deals with the analysis of data directly by marking the missing data as such and incorporating the missing data by defining new block types defined by the missing data in them. All three approaches are suggestions for tackling the third open problem set arising from the need to consider ‘bad’ data seriously – far

more seriously than the current cavalier practice of assuming network data are all ‘good’ when positional approaches are used.

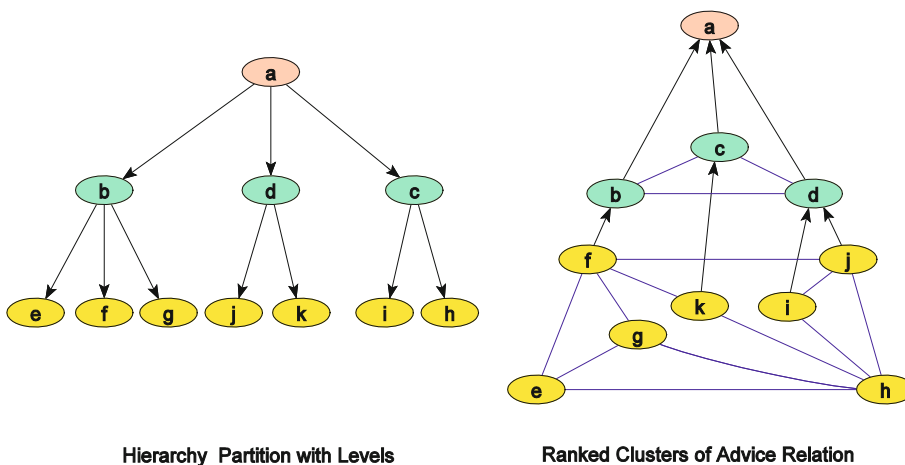
Most of the examples used in [14] feature binary networks and there is a clear need to extend generalized blockmodeling to consider valued networks, the fourth set of open problems. The balance theoretic partitioning can deal with valued signed network – as the example discussed above demonstrates. In one respect, conventional blockmodeling does this because many of the (dis)similarity measures that are used are applicable for valued networks. But for generalized blockmodeling this is a real problem. A crude way of dealing with valued networks is to impose some threshold to convert a valued network into one (or more) binary networks. The work in [22] suggests that this strategy is fraught with hazard and he has introduced a potentially fruitful approach for a generalized blockmodeling approach to valued network data.

A fifth open problem set is ushered in by the criterion function used to fit generalized blockmodels. For a given set of block types and a specific blockmodel type, we can say that the optimization approach provides empirical blockmodels that fit the best – subject to the caveat the relocation method is a local optimization procedure. If the criterion function is well defined (i. e. it is compatible with and sensitive to the equivalence used), then identifying partitions minimizing it all but guarantees the claim. And if the values of the criterion function are small (as in the examples used above) this claim is strengthened. However, this does not address the issue of how large a criterion function has to be to say that the specified block-

model does not fit the data well even though the partition returns the lowest possible value of the criterion function *for the data at hand*. The ‘best’ may not be good enough. Some statistical theory is needed for fully testing the fit of blockmodels to data in a principled fashion. Assembling this constitutes a solution set for the fifth set of open problems.

Dealing with multiple relations constitutes the sixth open problem set discussed here. This is acute for both the conventional and generalized approaches to the blockmodeling problem even though, at face value, it seems more acute for generalized blockmodeling. Conventional blockmodeling handles this issue by stacking the relations and computing the (dis)similarity measure across the full set of relations and mobilizing a clustering algorithm. However, this strategy implies that every relationship has the same set of block types for each relation. The example shown in Fig. 11 demonstrates the serious drawback with this assumption.

Two relations are shown. On the left is a hierarchical relation for a hypothetical bureaucratic structure, similar to the one in Fig. 1. The relation shown is ‘has authority over’ and a regular equivalence is shown with the levels that are color coded according to the partition. On the right is a hypothetical social relationship of advice seeking. Suppose, further, in this hypothetical organization that there are definite norms about advice seeking. People at the same bureaucratic level are free to seek advice from each other and it is acceptable for lower ranked subordinates to see advice from their bosses. In addition, suppose that the informal norms are such that bosses seeking advice from their subordinates is unacceptable and they



Hierarchy Partition with Levels

Ranked Clusters of Advice Relation

Positional Analysis and Blockmodeling, Figure 11

Two relations with the same partition but different blockmodels



never do so (or admit to doing so). The data on the right hand network are fully consistent with a ranked clusters model [12]. Within levels of the hierarchy, advice seeking, when it occurs, is fully symmetrical and the only directed ties are from subordinates at the lowest level to some of their bosses at the middle level and from the middle level to the top company official. This ranked clusters partition is exact and the clusters are identical for each relation in Fig. 11. The problem with the ‘stacking’ of relations approach is clear for this example. The blockmodels for each relationship are totally different even though the vertex partition for each relation is the same. Assuming, for example, that a single equivalence holds for both relations and computing some (dis)similarity will simply confound the two types of blockmodel with the result that the actual partitions will not be identified and, even if they were, they would be misrepresented in terms of the predicates of the blockmodel.

Analyzing multiple relations where the different relations can have quite different blockmodels is a deep problem, one that cannot be solved by the use of (dis)similarity measures as if they apply to each relation in the same fashion.

The seventh, and final, open problem set considered here has to do with the notion of evolving social networks. One of the basic assumptions of the positional approach, with blockmodeling as the primary strategy, is that there is a large and/or complex social network that we observe and the primary task is understand the structure of the network. This is done by delineating and describing the simpler blockmodel image. Another part of the basic assumption is that the underlying blockmodel image is the fundamental structure and the observed (surface) structure is an indicator only of the fundamental structure. There is a hint in the Supreme Court example that the voting structure did change over time. Of course, this could be due to the nature of the cases that happened to be considered in each of the terms studied. But even if there is some systematic evolution of the structure, the images depicted in Figs. 8 through 10 are no more than a description of change and does not deal with evolution as a structural process. The (deep) problem is to formulate models of network evolution for the fundamental structure(s) represented by the blockmodel while all of the data depict the observed surface model.

Blockmodeling methods, whether in the conventional or generalized mode, form a powerful set of tools for modeling, representing and understanding the structure of social networks. The many benefits stemming from the use of these tools, especially in the generalized blockmodeling variant, inspire confidence in this approach and suggest an

enormous potential for using these tools. However, there are many serious open problem sets that need to be solved before this potential is fully realized.

## Bibliography

1. Batagelj V (1997) Notes on blockmodeling. *Soc Netw* 19: 143–53
2. Batagelj V, Mrvar A (1998) Pajek – A Program for large network analysis. *Connections* 21:47–57
3. Batagelj V, Doreian P, Ferligoj A (1992) An optimization approach to regular equivalence. *Soc Netw* 14:121–35
4. Borgatti SP, Everett MG (1989) The class of all regular equivalences: Algebraic structure and computation. *Soc Netw* 11:159–72
5. Borgatti SP, Everett MG (2002) Notions of position in social network analysis. In: Marsden PV (ed) *Sociological Methodology*. Am Soc Assoc, Washington DC, pp 1–36
6. Borgatti SP, Everett MG, Freeman LC (2002) *Ucinet for Windows: Software for Social Network Analysis*, Analytic Technologies. Harvard, Massachusetts
7. Breiger RL, Boorman SA, Arabie P (1975) An algorithm for clustering relational data with applications to social network analysis and comparison to multidimensional scaling. *J Math Psychol* 12:328–83
8. Burt RS (1976) Positions in networks. *Soc Forces* 55:93–122
9. Cartwright D, Harary F (1956) Structural balance: A generalization of Heider’s theory. *Psychol Rev* 63:277–292
10. Davis JA (1967) Clustering and structural balance in graphs. *Hum Relat* 20:181–7
11. Doreian P, Mrvar A (1996) A partitioning approach to structural balance. *Soc Netw* 18:149–68
12. Doreian P, Batagelj V, Ferligoj A (2000) Symmetric-acyclic decomposition of networks. *J Classif* 17:3–28
13. Doreian P, Batagelj V, Ferligoj A (2005) Positional analysis of sociometric data. In: Carrington PJ, Scott J, Wasserman S (eds) *Models and Methods in Social Network Analysis*. Cambridge University Press, New York
14. Doreian P, Batagelj V, Ferligoj A (2005) *Generalized Blockmodeling*. Cambridge University Press, New York
15. Gold M, Doreian P, Taylor EF (2008) Understanding a collaborative effort to reduce racial and ethnic disparities in health care: Contributions from social network analysis. *Soc Sci Med* 67:1018–1027
16. Greenhouse L (2001, 2003, 2004) (a) In year of Florida vote, Supreme Court did much other work. *New York Times*, July 2; (b) In momentous term, justices remake the law and the court. *New York Times*, July 1; (c) The year Rehnquist may have lost his court. *New York Times*, July 4
17. Heider F (1946) Attitudes and cognitive organization. *J Psychol* 21:107–12
18. Lemann TB, Solomon RL (1952) Group Characteristics as Revealed in Sociometric Patterns and Personality Ratings. *Sociometry Monographs*, vol 27. Beacon House, Boston
19. Lorrain FP, White H (1971) Structural equivalence of individuals in networks. *J Math Soc* 1:49–80
20. White DR (1984) REGGE: A regular graph equivalence algorithm for computing role distances prior to blockmodeling. University of California, Irvine (Unpublished manuscript)



21. White DR, Reitz KP (1983) Graph and semigroup homomorphisms on networks and relations. *Soc Netw* 5:193–235
22. Žiberna A (2005) Generalized blockmodeling of valued networks. *Soc Netw* 29:105–26

## Possibility Theory

DIDIER DUBOIS, HENRI PRADE  
IRIT-CNRS, Université Paul Sabatier,  
Toulouse Cedex, France

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Historical Background  
Basic Notions of Possibility Theory  
Qualitative Possibility Theory  
Quantitative Possibility Theory  
Probability-Possibility Transformations  
Applications and Future Directions  
Bibliography

### Glossary

**Possibility distribution** A possibility distribution restricts a set of possible values for a variable of interest in an elastic way. It is represented by a mapping from a universe gathering the potential values of the variable to a scale such as the unit interval of the real line, or a finite linearly ordered set, expressing to what extent each value is possible for the variable. Thus, a possibility distribution restricts a set of more or less possible values belonging to a universe that may be also ordered such as a subpart of real line for a numerical variable, or not ordered if for instance the variable takes its value in the set of interpretations of a logical language. This may be used for representing uncertainty if the restriction pertains to possible values for an ill-known state of the world, or for representing preferences if the restriction encodes a set of values that are considered as more or less satisfactory for some purpose.

**Possibility measure** A possibility measure is a set function (increasing in the wide sense) that returns the maximum of a possibility distribution over a subset representing an event.

**Necessity measure** A necessity measure is a set function, associated by duality to a possibility measure through a relation expressing that an event is all the more necessarily true (all the more certain) as the opposite event is less possible. A necessity measure estimates to what

extent the information represented by the underlying possibility distribution entails the occurrence of the event.

**Guaranteed possibility** A guaranteed possibility measure is a set function (decreasing in the wide sense) that returns the minimum of a possibility distribution over a subset representing an event. While possibility measures evaluate the consistency of the information between an event and the available information represented by the underlying possibility distribution, guaranteed possibility measures capture another view of the idea of possibility related to the idea of (guaranteed) feasibility, or sufficiency condition.

**Possibilistic logic** Standard possibilistic logic is a weighted logic where formulas are pairs made of a classical logical formula and a weight that acts as a lower bound of the necessity of the logical formula. Extended possibilistic logics may include formulas weighted in terms of lower bounds of possibility or guaranteed possibility measures.

### Definition of the Subject

Possibility theory is the simplest uncertainty theory devoted to the modeling of incomplete information. It is characterized by the use of two basic dual set functions that respectively grade the possibility and the necessity of events. Possibility theory lies at the crossroads between fuzzy sets, probability and non-monotonic reasoning. Possibility theory is closely related to fuzzy sets if one considers that a possibility distribution is a particular fuzzy set (of mutually exclusive) possible values. However fuzzy sets and fuzzy logic are primarily motivated by the representation of gradual properties while possibility theory handles the uncertainty of classical (or fuzzy) propositions. Possibility theory can be cast either in an ordinal or in a numerical setting. Qualitative possibility theory is closely related to belief revision theory, and common-sense reasoning with exception-tainted knowledge in Artificial Intelligence. It has been axiomatically justified in a decision-theoretic framework in the style of Savage, thus providing a foundation for qualitative decision theory. Quantitative possibility theory is the simplest framework for statistical reasoning with imprecise probabilities. As such it has close connections with random set theory and confidence intervals, and can provide a tool for uncertainty propagation with limited statistical or subjective information.

### Introduction

Possibility theory is an uncertainty theory devoted to the handling of incomplete information. To a large extent, it

is similar to probability theory because it is based on set-functions. It differs from the latter by the use of a pair of dual set functions (possibility and necessity measures) instead of only one. Besides, it is not additive and makes sense on ordinal structures. The name “Theory of Possibility” was coined by Zadeh [1], who was inspired by a paper by Gaines and Kohout [2]. In Zadeh’s view, possibility distributions were meant to provide a graded semantics to natural language statements. However, possibility and necessity measures can also be the basis of a full-fledged representation of partial belief that parallels probability. It can be seen either as a coarse, non-numerical version of probability theory, or a framework for reasoning with extreme probabilities, or yet a simple approach to reasoning with imprecise probabilities [3].

After reviewing pioneering contributions to possibility theory, we recall its basic concepts and present the two main directions along which it has developed: the qualitative and quantitative settings. Both approaches share the same basic “maxitivity” axiom. They differ when it comes to conditioning, and to independence notions.

### Historical Background

Zadeh was not the first scientist to speak about formalizing notions of possibility. The modalities *possible* and *necessary* have been used in philosophy at least since the Middle-Ages in Europe, based on Aristotle’s works. More recently they became the building blocks of Modal Logics that emerged at the end of the first decade of the XXth century from the works of C.I. Lewis (see Hughes and Cresswell [5]). In this approach, possibility and necessity are all-or-nothing notions, and handled at the syntactic level. More recently, and independently from Zadeh’s view, the notion of possibility, as opposed to probability, was central in the works of one economist, Shackle, and in those of two philosophers, D. Lewis and L.J. Cohen.

#### G.L.S. Shackle

A graded notion of possibility was introduced as a full-fledged approach to uncertainty and decision in the 1940–1970’s by the English economist G.L.S. Shackle [6], who called *degree of potential surprise* of an event its degree of impossibility, that is, the degree of necessity of the opposite event. Shackle’s notion of possibility is basically epistemic, it is a “character of the chooser’s particular state of knowledge in his present”. Impossibility is understood as disbelief. Potential surprise is valued on a disbelief scale, namely a positive interval of the form  $[0, y^*]$ , where  $y^*$  denotes the absolute rejection of the event to which it is

assigned. In case everything is possible, all mutually exclusive hypotheses have zero surprise. At least one elementary hypothesis must carry zero potential surprise. The degree of surprise of an event, a set of elementary hypotheses, is the degree of surprise of its least surprising realization. The disbelief notion introduced later by Spohn [7] employs the same type of convention as potential surprise, but using the set of natural integers as a disbelief scale. Shackle also introduces a notion of conditional possibility, whereby the degree of surprise of a conjunction of two events  $A$  and  $B$  is equal to the maximum of the degree of surprise of  $A$ , and of the degree of surprise of  $B$ , should  $A$  prove true.

#### D. Lewis

In his 1973 book [8] the philosopher David Lewis considers a graded notion of possibility in the form of a relation between possible worlds he calls *comparative possibility*. He equates this concept of possibility to a notion of similarity between possible worlds. This non-symmetric notion of similarity is also comparative, and is meant to express statements of the form: *a world  $j$  is at least as similar to world  $i$  as world  $k$  is*. Comparative similarity of  $j$  and  $k$  with respect to  $i$  is interpreted as the comparative possibility of  $j$  with respect to  $k$  viewed from world  $i$ . Such relations are assumed to be complete pre-orderings and are instrumental in defining the truth conditions of counterfactual statements. Comparative possibility relations  $\geq_{\Pi}$  obey the key axiom: for all events  $A, B, C$ ,

$$A \geq_{\Pi} B \text{ implies } C \cup A \geq_{\Pi} C \cup B.$$

This axiom was later independently proposed by the first author [9] in an attempt to derive a possibilistic counterpart to comparative probabilities. Interestingly, the connection between numerical possibility and similarity is currently investigated by Sudkamp [10].

#### L.J. Cohen

A framework very similar to the one of Shackle was proposed by the philosopher L.J. Cohen [11] who considered the problem of legal reasoning. He introduced so-called *Baconian probabilities* understood as degrees of provability. The idea is that it is hard to prove someone guilty at the court of law by means of pure statistical arguments. The basic feature of degrees of provability is that a hypothesis and its negation cannot both be provable together to any extent (the contrary being a case for inconsistency). Such degrees of provability coincide with necessity measures.

### L.A. Zadeh

In his seminal paper [1] Zadeh proposed an interpretation of membership functions of fuzzy sets as possibility distributions encoding flexible constraints induced by natural language statements. Zadeh articulated the relationship between possibility and probability, noticing that what is probable must preliminarily be possible. However, the view of possibility degrees developed in his paper refers to the idea of graded feasibility (degrees of ease, as in the example of “how many eggs can Hans eat for his breakfast”) rather than to the epistemic notion of plausibility laid bare by Shackle. Nevertheless, the key axiom of “maxitivity” for possibility measures is highlighted. In two subsequent articles [12,13], Zadeh acknowledged the connection between possibility theory, belief functions and upper/lower probabilities, and proposed their extensions to fuzzy events and fuzzy information granules.

### Basic Notions of Possibility Theory

The basic building blocks of possibility theory were first extensively described in the authors’ books [14,15] (see also [16]). Let  $S$  be a set of states of affairs (or descriptions thereof), or states for short. A possibility distribution is a mapping  $\pi$  from  $S$  to a totally ordered scale  $L$ , with top 1 and bottom 0, such as the unit interval. The function  $\pi$  represents the state of knowledge of an agent (about the actual state of affairs) distinguishing what is plausible from what is less plausible, what is the normal course of things from what is not, what is surprising from what is expected. It represents a flexible restriction on what is the actual state with the following conventions (similar to probability, but opposite to Shackle’s potential surprise scale):

- $\pi(s) = 0$  means that state  $s$  is rejected as impossible;
- $\pi(s) = 1$  means that state  $s$  is totally possible (= plausible).

If  $S$  is exhaustive, at least one of the elements of  $S$  should be the actual world, so that  $\exists s, \pi(s) = 1$  (normalization). Distinct values may simultaneously have a degree of possibility equal to 1.

Possibility theory is driven by the principle of minimal specificity. It states that any hypothesis not known to be impossible cannot be ruled out. A possibility distribution  $\pi$  is said to be at least as specific as another  $\pi'$  if and only if for each state of affairs  $s$ :  $\pi(s) \leq \pi'(s)$  (Yager [17]). Then,  $\pi$  is at least as restrictive and informative as  $\pi'$ .

In the possibilistic framework, extreme forms of partial knowledge can be captured, namely:

- Complete knowledge: for some  $s_0$ ,  $\pi(s_0) = 1$  and  $\pi(s) = 0, \forall s \neq s_0$  (only  $s_0$  is possible)

- Complete ignorance:  $\pi(s) = 1, \forall s \in S$ , (all states are possible).

Given a simple query of the form “does event  $A$  occur?” where  $A$  is a subset of states, the response to the query can be obtained by computing degrees of possibility and necessity, respectively (if the possibility scale  $L = [0, 1]$ ):

$$\Pi(A) = \sup_{s \in A} \pi(s); N(A) = \inf_{s \notin A} 1 - \pi(s).$$

$\Pi(A)$  evaluates to what extent  $A$  is consistent with  $\pi$ , while  $N(A)$  evaluates to what extent  $A$  is certainly implied by  $\pi$ . The possibility-necessity duality is expressed by  $N(A) = 1 - \Pi(A^c)$ , where  $A^c$  is the complement of  $A$ . Generally,  $\Pi(S) = N(S) = 1$  and  $\Pi(\emptyset) = N(\emptyset) = 0$ . Possibility measures satisfy the basic “maxitivity” property  $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$ . Necessity measures satisfy an axiom dual to that of possibility measures, namely  $N(A \cap B) = \min(N(A), N(B))$ . On infinite spaces, these axioms must hold for infinite families of sets.

Human knowledge is often expressed in a declarative way using statements to which belief degrees are attached. It corresponds to expressing constraints the world is supposed to comply with. Certainty-qualified pieces of uncertain information of the form “ $A$  is certain to degree  $\alpha$ ” can then be modelled by the constraint  $N(A) \geq \alpha$ . The least specific possibility distribution reflecting this information is [15]:

$$\pi_{(A,\alpha)}(s) = \begin{cases} 1, & \text{if } s \in A \\ 1 - \alpha, & \text{otherwise} \end{cases} \quad (1)$$

Acquiring further pieces of knowledge leads to updating  $\pi_{(A,\alpha)}$  into some  $\pi < \pi_{(A,\alpha)}$ .

Apart from  $\Pi$  and  $N$ , a measure of *guaranteed possibility* or *sufficiency* can be defined [18]:  $\Delta(A) = \inf_{s \in A} \pi(s)$ . In contrast,  $\Pi$  appears to be a measure of *potential possibility*. It estimates to what extent *all* states in  $A$  are actually possible according to evidence.  $\Delta(A)$  can be used as a degree of evidential support for  $A$ . Uncertain statements of the form “ $A$  is possible to degree  $\beta$ ” often mean that all realizations of  $A$  are possible to degree  $\beta$ . They can then be modelled by the constraint  $\Delta(A) \geq \beta$ . It corresponds to the idea of observed evidence. This type of information is better exploited by assuming an informational principle opposite to the one of minimal specificity, namely, any situation not yet observed is tentatively considered as impossible. This is similar to closed-world assumption. The most specific distribution  $\delta_{(A,\beta)}$  in agreement with  $\Delta(A) \geq \beta$  is:

$$\delta_{(A,\beta)}(s) = \begin{cases} \beta, & \text{if } s \in A \\ 0, & \text{otherwise} \end{cases}$$

Acquiring further pieces of evidence leads to updating  $\delta_{(A,\beta)}$  into some wider distribution  $\delta > \delta_{(A,\beta)}$ . Such evidential support functions do not behave with the same conventions as possibility distributions:  $\delta(s) = 1$  means that  $S$  is guaranteed to be possible, because of a high evidential support, while  $\delta(s) = 0$  only means that  $S$  has not been observed yet (hence is of unknown possibility). Distributions  $\delta$  are generally not normalized to 1, and serve as lower bounds to possibility distributions  $\pi$  (because what is observed must be possible). Such a bipolar representation of information using pairs  $(\delta, \pi)$  may provide a natural interpretation of interval-valued fuzzy sets. Note that possibility distributions induced from certainty-qualified pieces of knowledge combine conjunctively, by discarding possible states, while evidential support distributions induced by possibility-qualified pieces of evidence combine disjunctively, by accumulating possible states.

Notions of conditioning and independence were studied for possibility measures. Conditional possibility is defined similarly to probability theory using a Bayesian like equation of the form [15]

$$\Pi(B \cap A) = \Pi(B|A) \star \Pi(A).$$

However, in the ordinal setting the operation  $\star$  cannot be a product and is changed into the minimum. In the numerical setting, there are several ways to define conditioning, not all of which have this form [19]. There are several variants of possibilistic independence [20,21,22]. Generally, independence in possibility theory is neither symmetric, nor insensitive to negation. For non Boolean variables, independence between events is not equivalent to independence between variables. Joint possibility distributions on Cartesian products of domains can be represented by means of graphical structures similar to Bayesian networks for joint probabilities (see [23,24]. Such graphical structures can be taken advantage of for evidence propagation [25] or learning [26].

### Qualitative Possibility Theory

This section is restricted to the case of a finite state space  $S$ , supposed to be the set of interpretations of a formal propositional language. In other words,  $S$  is the universe induced by Boolean attributes. A plausibility ordering is a complete pre-order of states denoted by  $\geq_\pi$ , which induces a well-ordered partition  $\{E_1, \dots, E_n\}$  of  $S$ . It is the comparative counterpart of a possibility distribution  $\pi$ , i. e.,  $s \geq_\pi s'$  if and only if  $\pi(s) \geq \pi(s')$ . Indeed it is more natural to expect that an agent will supply ordinal rather than numerical information about his beliefs. By convention  $E_1$  contains the most normal states of fact,  $E_n$  the least plausible,

or most surprising ones. Denoting  $\text{argmax}(A)$  any most plausible state  $s_0 \in A$ , ordinal counterparts of possibility and necessity measures [9] are then defined as follows:  $\{s\} \geq_\Pi \emptyset$  for all  $s \in S$  and

$$\begin{aligned} A \geq_\Pi B & \text{ if and only if } \max(A) \geq_\pi \max(B) \\ A \geq_N B & \text{ if and only if } \max(B^c) \geq_\pi \max(A^c). \end{aligned}$$

Possibility relations  $\geq_\Pi$  are those of Lewis and satisfy the characteristic property

$$A \geq_\Pi B \text{ implies } C \cup A \geq_\Pi C \cup B$$

while necessity relations can also be defined as  $A \geq_N B$  if and only if  $B^c \geq_\Pi A^c$ , and satisfy a similar axiom:

$$A \geq_N B \text{ implies } C \cap A \geq_N C \cap B.$$

The latter coincide with epistemic entrenchment relations in the sense of belief revision theory [27,28]. Conditioning a possibility relation  $\geq_\Pi$  by an non-impossible event  $C >_\Pi \emptyset$  means deriving a relation  $\geq_\Pi^C$  such that

$$A \geq_\Pi^C B \text{ if and only if } A \cap C \geq_\Pi B \cap C.$$

The notion of independence for comparative possibility theory was studied in Dubois et al. [22], for independence between events, and Ben Amor et al. [29] between variables.

### Non-monotonic Inference

Suppose  $S$  is equipped with a plausibility ordering. The main idea behind qualitative possibility theory is that the state of the world is always believed to be as normal as possible, neglecting less normal states.  $A \geq_\Pi B$  really means that there is a normal state where  $A$  holds that is at least as normal as any normal state where  $B$  holds. The dual case  $A \geq_N B$  is intuitively understood as “ $A$  is at least as certain as  $B$ ”, in the sense that there are states where  $B$  fails to hold that are at least as normal as the most normal state where  $A$  does not hold. In particular, the events accepted as true are those which are true in all the most plausible states, namely the ones such that  $A >_N \emptyset$ . These assumptions lead us to interpret the plausible inference  $A| \approx B$  of a proposition  $B$  from another  $A$ , under a state of knowledge  $\geq_\Pi$  as follows:  *$B$  should be true in all the most normal states where  $A$  is true*, which means  $B >_\Pi^A B^c$  in terms of ordinal conditioning, that is,  $A \cap B$  is more plausible than  $A \cap B^c$ .  $A| \approx B$  also means that the agent considers  $B$  as an accepted belief in the context  $A$ .

This kind of inference is non-monotonic in the sense that  $A| \approx B$  does not always imply  $A \cap C| \approx B$  for any

additional information  $C$ . This is similar to the fact that a conditional probability  $P(B|A \cap C)$  may be low even if  $P(B|A)$  is high. The properties of the consequence relation  $|\approx$  are now well-understood, and are precisely the ones laid bare by Lehmann and Magidor [30] for their so-called “rational inference”. Monotonicity is only partially restored:  $A|\approx B$  implies  $A \cap C|\approx B$  holds provided that  $A|\approx C^c$  does not hold (i. e. that states where  $A$  is true do not typically violate  $C$ ). This property is called *rational monotony*, and, along with some more standard ones (like closure under conjunction), characterizes default possibilistic inference  $|\approx$ . In fact, the set  $\{B, A|\approx B\}$  of accepted beliefs in the context  $A$  is deductively closed, which corresponds to the idea that the agent reasons with accepted beliefs in each context as if they were true, until some event occurs that modifies this context. This closure property is enough to justify a possibilistic approach [31] and adding the rational monotonicity property ensures the existence of a single possibility relation generating the consequence relation  $|\approx$  [32].

Rather than being constructed from scratch, plausibility orderings can be generated by a set of if-then rules tainted with unspecified exceptions. This set forms a knowledge base supplied by an agent. Each rule “if  $A$  then  $B$ ” is understood as a constraint of the form  $A \cap B >_{\Pi} A \cap B^c$  on possibility relations. There exists a single minimally specific element in the set of possibility relations satisfying all constraints induced by rules (unless the latter are inconsistent). It corresponds to the most compact plausibility ranking of states induced by the rules [32]. This ranking can be computed by an algorithm originally proposed by Pearl [33].

### Possibilistic Logic

Qualitative possibility relations can be represented by (and only by) possibility measures ranging on any totally ordered set  $L$  (especially a finite one) [9]. This absolute representation on an ordinal scale is slightly more expressive than the purely relational one. When the finite set  $S$  is large and generated by a propositional language, qualitative possibility distributions can be efficiently encoded in possibilistic logic [34]. A possibilistic logic base  $K$  is a set of pairs  $(\phi, \alpha)$ , where  $\phi$  is a Boolean expression and  $\alpha$  is an element of  $L$ . This pair encodes the constraint  $N(\phi) \geq \alpha$  where  $N(\phi)$  is the degree of necessity of the set of models of  $\phi$ . Each prioritized formula  $(\phi, \alpha)$  has a fuzzy set of models (described in Sect. “Basic Notions of Possibility Theory”) and the fuzzy intersection of the fuzzy sets of models of all prioritized formulas in  $K$  yields the associated plausibility ordering on  $S$ .

Syntactic deduction from a set of prioritized clauses is achieved by refutation using an extension of the standard resolution rule, whereby  $(\phi \vee \psi, \min(\alpha, \beta))$  can be derived from  $(\phi \vee \xi, \alpha)$  and  $(\psi \vee \neg\xi, \beta)$ . This rule, which evaluates the validity of an inferred proposition by the validity of the weakest premiss, goes back to Theophrastus, a disciple of Aristotle. Possibilistic logic is an inconsistency-tolerant extension of propositional logic that provides a natural semantic setting for mechanizing non-monotonic reasoning [35], with a computational complexity close to that of propositional logic. See [36] for a detailed introduction to possibilistic logic, its syntactic and semantic aspects, different extensions that involve time, for instance, or that encode constraints on lower bounds of possibility or guaranteed possibility measures, or that perform multiple source information fusion. See [37] for a sketch of further extensions for handling groups of agents’ beliefs and mutual beliefs.

Another compact representation of qualitative possibility distributions is the possibilistic directed graph, which uses the same conventions as Bayesian nets, but relies on an ordinal notion of conditional possibility [15]

$$\Pi(B|A) = \begin{cases} 1, & \text{if } \Pi(B \cap A) = \Pi(A) \\ \Pi(B \cap A), & \text{otherwise.} \end{cases}$$

Joint possibility distributions can be decomposed into a conjunction of conditional possibility distributions (using minimum) in a way similar to Bayes nets [38]. It is based on a symmetric notion of qualitative independence  $\Pi(B \cap A) = \min(\Pi(A), \Pi(B))$  that is weaker than the causal-like condition  $\Pi(B|A) = \Pi(B)$  [22]. Ben Amor and Benferhat [39] investigate the properties of qualitative independence that enable local inferences to be performed in possibilistic nets.

### Decision-Theoretic Foundations

Zadeh [1] hinted that “since our intuition concerning the behavior of possibilities is not very reliable”, our understanding of them “would be enhanced by the development of an axiomatic approach to the definition of subjective possibilities in the spirit of axiomatic approaches to the definition of subjective probabilities”. Decision-theoretic justifications of qualitative possibility were recently devised, in the style of Savage [40]. On top of the set of states, assume there is a set  $X$  of consequences of decisions. A decision, or act, is modelled as a mapping  $f$  from  $S$  to  $X$  assigning to each state  $S$  its consequence  $f(s)$ . The axiomatic approach consists in proposing properties of a preference relation  $\succeq$  between acts so that a representation of this



relation by means of a preference functional  $W(f)$  is ensured, that is, act  $f$  is as good as act  $g$  (denoted  $f \succeq g$ ) if and only if  $W(f) \geq W(g)$ .  $W(f)$  depends on the agent's knowledge about the state of affairs, here supposed to be a possibility distribution  $\pi$  on  $S$ , and the agent's goal, modelled by a utility function  $u$  on  $X$ . Both the utility function and the possibility distribution map to the same finite chain  $L$ . A pessimistic criterion  $W_{\pi}^{-}(f)$  is of the form:

$$W_{\pi}^{-}(f) = \min_{s \in S} \max(n(\pi(s)), u(f(s)))$$

where  $n$  is the order-reversing map of  $L$ .  $n(\pi(s))$  is the degree of certainty that the state is not  $s$  (hence the degree of surprise of observing  $s$ ),  $u(f(s))$  the utility of choosing act  $f$  in state  $s$ .  $W_{\pi}^{-}(f)$  is all the higher as all states are either very surprising or have high utility. This criterion is actually a prioritized extension of the Wald maximin criterion. The latter is recovered if  $\pi(s) = 1$  (top of  $L$ )  $\forall s \in S$ . According to the pessimistic criterion, acts are chosen according to their worst consequences, restricted to the most plausible states  $S^* = \{s, \pi(s) \geq n(W_{\pi}^{-}(f))\}$ . The optimistic counterpart of this criterion is:

$$W_{\pi}^{+}(f) = \max_{s \in S} \min(\pi(s), u(f(s)))$$

$W_{\pi}^{+}(f)$  is all the higher as there is a very plausible state with high utility. The optimistic criterion was first proposed by Yager [41] and the pessimistic criterion by Whalen [42]. These optimistic and pessimistic possibilistic criteria are particular cases of a more general criterion based on the Sugeno integral [43] specialized to possibility and necessity of fuzzy events [1,14]:

$$S_{\gamma,u}(f) = \max_{\lambda \in L} \min(\lambda, \gamma(F_{\lambda}))$$

where  $F_{\lambda} = \{s \in S, u(f(s)) \geq \lambda\}$ ,  $\gamma$  is a monotonic set function that reflects the decision-maker attitude in front of uncertainty:  $\gamma(A)$  is the degree of confidence in event  $A$ . If  $\gamma = \Pi$ , then  $S_{\Pi,u}(f) = W_{\pi}^{+}(f)$ . Similarly, if  $\gamma = N$ , then  $S_{N,u}(f) = W_{\pi}^{-}(f)$ .

For any acts  $f, g$ , and any event  $A$ , let  $fAg$  denote an act consisting of choosing  $f$  if  $A$  occurs and  $g$  if its complement occurs. Let  $f \wedge g$  (resp.  $f \vee g$ ) be the act whose results yield the worst (resp. best) consequence of the two acts in each state. Constant acts are those whose consequence is fixed regardless of the state. A result in [44,45] provides an act-driven axiomatization of these criteria, and enforces possibility theory as a "rational" representation of uncertainty for a finite state space  $S$ :

**Theorem 1** Suppose the preference relation  $\succeq$  on acts obeys the following properties:

1.  $(X^S, \succeq)$  is a complete preorder.
2. There are two acts such that  $f \succ g$ .
3.  $\forall A, \forall g$  and  $h$  constant,  $\forall f, g \succeq h$  implies  $gAf \succeq hAf$ .
4. If  $f$  is constant,  $f \succ h$  and  $g \succ h$  imply  $f \wedge g \succ h$ .
5. If  $f$  is constant,  $h \succ f$  and  $h \succ g$  imply  $h \succ f \vee g$ .

Then there exists a finite chain  $L$ , an  $L$ -valued monotonic set-function  $\gamma$  on  $S$  and an  $L$ -valued utility function  $u$ , such that  $\succeq$  is representable by a Sugeno integral of  $u(f)$  with respect to  $\gamma$ . Moreover  $\gamma$  is a necessity (resp. possibility) measure as soon as property (iv) (resp. (v)) holds for all acts. The preference functional is then  $W_{\pi}^{-}(f)$  (resp.  $W_{\pi}^{+}(f)$ ).

Axioms (4–5) contradict expected utility theory. They become reasonable if the value scale is finite, decisions are one-shot (no compensation) and provided that there is a big step between any level in the qualitative value scale and the adjacent ones. In other words the preference pattern  $f \succ h$  always means that  $f$  is significantly preferred to  $h$ , to the point of considering the value of  $h$  negligible in front of the value of  $f$ . The above result provides decision-theoretic foundations of possibility theory, whose axioms can thus be tested from observing the choice behavior of agents. See [46] for another approach to comparative possibility relations, more closely relying on Savage axioms but giving up any comparability between utility and plausibility levels. The drawback of these and other qualitative decision criteria is their lack of discrimination power [47]. To overcome it, refinements of possibilistic criteria were recently proposed, based on lexicographic schemes [48]. These new criteria turn out to be representable by a classical (but big-stepped) expected utility criterion.

## Quantitative Possibility Theory

The phrase "quantitative possibility" refers to the case when possibility degrees range in the unit interval. In that case, a precise articulation between possibility and probability theories is useful to provide an interpretation to possibility and necessity degrees. Several such interpretations can be consistently devised. See [49] for a detailed survey. A degree of possibility can be viewed as an upper probability bound [50], and a possibility distribution can be viewed as a likelihood function [51]. A possibility measure is also a special case of a Shafer plausibility function [52]. Following a very different approach, possibility theory can account for probability distributions with extreme values, infinitesimal [7] or having big steps [53]. There are finally close connections between possibility theory and idempotent analysis [54,55]. The theory of large deviations in probability theory [56] also handles set-functions that look

like possibility measures [57]. Here we focus on the role of possibility theory in the theory of imprecise probability.

### Possibility as Upper Probability

Let  $\pi$  be a possibility distribution where  $\pi(s) \in [0, 1]$ . Let  $\mathbf{P}(\pi)$  be the set of probability measures  $P$  such that  $P \leq \Pi$ , i.e.  $\forall A \subseteq S, P(A) \leq \Pi(A)$ . Then the possibility measure  $\Pi$  coincides with the upper probability function  $P^*$  such that  $P^*(A) = \sup\{P(A), P \in \mathbf{P}(\pi)\}$  while the necessity measure  $N$  is the lower probability function  $P_*$  such that  $P_*(A) = \inf\{P(A), P \in \mathbf{P}(\pi)\}$ ; see [50,58] for details.  $P$  and  $\pi$  are said to be consistent if  $P \in \mathbf{P}(\pi)$ . The connection between possibility measures and imprecise probabilistic reasoning is especially promising for the efficient representation of non-parametric families of probability functions, and it makes sense even in the scope of modelling linguistic information [59].

A possibility measure can be computed from a set of nested confidence subsets  $\{A_1, A_2, \dots, A_m\}$  where  $A_i \subset A_{i+1}, i = 1 \dots m-1$ . Each confidence subset  $A_i$  is attached a positive confidence level  $\lambda_i$  interpreted as a lower bound of  $P(A_i)$ , hence a necessity degree. It is viewed as a certainty-qualified statement that generates a possibility distribution  $\pi_i$  according to Sect. “Basic Notions of Possibility Theory”. The corresponding possibility distribution is

$$\pi(s) = \min_{i=1, \dots, m} \pi_i(s) = \begin{cases} 1, & \text{if } u \in A_1 \\ 1 - \lambda_{j-1}, & \text{if } j = \max\{i : s \notin A_i\} > 1. \end{cases}$$

The information modelled by  $\pi$  can also be viewed as a nested random set  $\{(A_i, v_i), i = 1, \dots, m\}$ , where  $v_i = \lambda_i - \lambda_{i-1}$ . This framework allows for imprecision (reflected by the size of the  $A_i$ 's) and uncertainty (the  $v_i$ 's). And  $v_i$  is the probability that the agent only knows that  $A_i$  contains the actual state (it is not  $P(A_i)$ ). The random set view of possibility theory is well adapted to the idea of imprecise statistical data, as developed in [60,61]. Namely, given a bunch of imprecise (not necessarily nested) observations (called focal sets),  $\pi$  supplies an approximate representation of the data, as  $\pi(s) = \sum_{i: s \in A_i} v_i$ .

The set  $\mathbf{P}(\pi)$  contains many probability distributions, arguably too many. Neumaier [62] has recently proposed a related framework, in a different terminology, for representing smaller subsets of probability measures using two possibility distributions instead of one. He basically uses a pair of distributions  $(\delta, \pi)$  (in the sense of Sect. “Basic Notions of Possibility Theory”) of distributions, he calls “cloud”, where  $\delta$  is a guaranteed possibility distribution

(in our terminology) such that  $\pi \geq \delta$ . A cloud models the (generally non-empty) set  $\mathbf{P}(\pi) \cap \mathbf{P}(1 - \delta)$ , viewing  $1 - \delta$  as a standard possibility distribution.

### Conditioning

There are two kinds of conditioning that can be envisaged upon the arrival of new information  $E$ . The first method presupposes that the new information alters the possibility distribution  $\pi$  by declaring all states outside  $E$  impossible. The conditional measure  $\pi(\cdot|E)$  is such that  $\Pi(B|E) \cdot \Pi(E) = \Pi(B \cap E)$ . This is formally Dempster rule of conditioning of belief functions, specialized to possibility measures. The conditional possibility distribution representing the weighted set of confidence intervals is,

$$\pi(s|E) = \begin{cases} \frac{\pi(s)}{\Pi(E)}, & \text{if } s \in E \\ 0, & \text{otherwise.} \end{cases}$$

De Baets et al. [63] provide a mathematical justification of this notion in an infinite setting, as opposed to the min-based conditioning of qualitative possibility theory. Indeed, the maxitivity axiom extended to the infinite setting is not preserved by the min-based conditioning. The product-based conditioning leads to a notion of independence of the form  $\Pi(B \cap E) = \Pi(B) \cdot \Pi(E)$  whose properties are very similar to the ones of probabilistic independence [21].

Another form of conditioning [64,65], more in line with the Bayesian tradition, considers that the possibility distribution  $\pi$  encodes imprecise statistical information, and event  $E$  only reflects a feature of the current situation, not of the state in general. Then the value  $\Pi(B|E) = \sup\{P(B|E), P(E) > 0, P \leq \Pi\}$  is the result of performing a sensitivity analysis of the usual conditional probability over  $\mathbf{P}(\pi)$  (Walley [66]). Interestingly, the resulting set-function is again a possibility measure, with distribution

$$\pi(s||E) = \begin{cases} \max\left(\pi(s), \frac{\pi(s)}{\pi(s) + N(E)}\right), & \text{if } s \in E \\ 0, & \text{otherwise.} \end{cases}$$

It is generally less specific than  $\pi$  on  $E$ , as clear from the above expression, and becomes non-informative when  $N(E) = 0$  (i.e. if there is no information about  $E$ ). This is because  $\pi(\cdot||E)$  is obtained from the focusing of the generic information  $\pi$  over the reference class  $E$ . On the contrary,  $\pi(\cdot|E)$  operates a revision process on  $\pi$  due to additional knowledge asserting that states outside  $E$  are impossible. See De Cooman [65] for a detailed study of this form of conditioning.

### Probability-Possibility Transformations

The problem of transforming a possibility distribution into a probability distribution and conversely is meaningful in the scope of uncertainty combination with heterogeneous sources (some supplying statistical data, other linguistic data, for instance). It is useful to cast all pieces of information in the same framework. The basic requirement is to respect the consistency principle  $\Pi \geq P$ . The problem is then either to pick a probability measure in  $\mathbf{P}(\pi)$ , or to construct a possibility measure dominating  $P$ .

There are two basic approaches to possibility/probability transformations, which both respect a form of probability-possibility consistency. One, due to Klir [67,68] is based on a principle of information invariance, the other [69] is based on optimizing information content. Klir assumes that possibilistic and probabilistic information measures are commensurate. Namely, the choice between possibility and probability is then a mere matter of translation between languages “neither of which is weaker or stronger than the other” (quoting Klir and Parviz [70]). It suggests that entropy and imprecision capture the same facet of uncertainty, albeit in different guises. The other approach, recalled here, considers that going from possibility to probability leads to increase the precision of the considered representation (as we go from a family of nested sets to a random element), while going the other way around means a loss of specificity.

#### From Possibility to Probability

The most basic example of transformation from possibility to probability is the Laplace principle of insufficient reason claiming that what is equally possible should be considered as equally probable. A generalized Laplacean indifference principle is then adopted in the general case of a possibility distribution  $\pi$ : the weights  $v_i$  bearing the sets  $A_i$  from the nested family of levels cuts of  $\pi$  are uniformly distributed on the elements of these cuts  $A_i$ . Let  $P_i$  be the uniform probability measure on  $A_i$ . The resulting probability measure is  $P = \sum_{i=1, \dots, m} v_i \cdot P_i$ . This transformation, already proposed in 1982 [71] comes down to selecting the center of gravity of the set  $\mathbf{P}(\pi)$  of probability distributions dominated by  $\pi$ . This transformation also coincides with Smets’ pignistic transformation [72] and with the Shapley value of the “unanimity game” (another name of the necessity measure) in game theory. The rationale behind this transformation is to minimize arbitrariness by preserving the symmetry properties of the representation. This transformation from possibility to probability is one-to-one. Note that the definition of this transformation does not use the nestedness property of cuts of

the possibility distribution. It applies all the same to non-nested random sets (or belief functions) defined by pairs  $\{(A_i, v_i), i = 1, \dots, m\}$ , where  $v_i$  are non-negative reals such that  $\sum_{i=1, \dots, m} v_i = 1$ .

#### From Objective Probability to Possibility

From probability to possibility, the rationale of the transformation is not the same according to whether the probability distribution we start with is subjective or objective [73]. In the case of a statistically induced probability distribution, the rationale is to preserve as much information as possible. This is in line with the handling of  $\Delta$ -qualified pieces of information representing observed evidence, considered in Sect. “Basic Notions of Possibility Theory”; hence we select as the result of the transformation of a probability measure  $P$ , the most specific possibility measure in the set of those dominating  $P$  [69]. This most specific element is generally unique if  $P$  induces a linear ordering on  $S$ . Suppose  $S$  is a finite set. The idea is to let  $\Pi(A) = P(A)$ , for these sets  $A$  having minimal probability among other sets having the same cardinality as  $A$ . If  $p_1 > p_2 > \dots > p_n$ , then  $\Pi(A) = P(A)$  for sets  $A$  of the form  $\{s_i, \dots, s_n\}$ , and the possibility distribution is defined as  $\pi_P(s_i) = \sum_{j=i, \dots, n} p_j$ . Note that  $\pi_P$  is a kind of cumulative distribution of  $P$ . If there are equiprobable elements, the unicity of the transformation is preserved if equiprobability of the corresponding elements is enforced. In this case it is a bijective transformation as well. Recently, this transformation was used to prove a rather surprising agreement between probabilistic indeterminateness as measured by Shannon entropy, and possibilistic non-specificity. Namely it is possible to compare probability measures on finite sets in terms of their relative *peakedness* (a concept adapted from Birnbaum [74]) by comparing the relative specificity of their possibilistic transforms. Namely let  $P$  and  $Q$  be two probability measures on  $S$  and  $\pi_P, \pi_Q$  the possibility distributions induced by our transformation. It can be proved that if  $\pi_P \geq \pi_Q$  (i. e.  $P$  is less peaked than  $Q$ ) then the Shannon entropy of  $P$  is higher than the one of  $Q$  [75]. This result give some grounds to the intuitions developed by Klir [67], without assuming any commensurability between entropy and specificity indices.

#### Possibility Distributions Induced by Prediction Intervals

In the continuous case, moving from objective probability to possibility means adopting a representation of uncertainty in terms of prediction intervals around the mode viewed as the “most frequent value”. Extracting a predic-

tion interval from a probability distribution or devising a probabilistic inequality can be viewed as moving from a probabilistic to a possibilistic representation. Namely suppose a non-atomic probability measure  $P$  on the real line, with unimodal density  $p$ , and suppose one wishes to represent it by an interval  $I$  with a prescribed level of confidence  $P(I) = \gamma$  of hitting it. The most natural choice is the most precise interval ensuring this level of confidence. It can be proved that this interval is of the form of a cut of the density, i.e.  $I_\gamma = \{s, p(s) \geq \theta\}$  for some threshold  $\theta$ . Moving the degree of confidence from 0 to 1 yields a nested family of prediction intervals that form a possibility distribution  $\pi$  consistent with  $P$ , the most specific one actually, having the same support and the same mode as  $P$  and defined by ([69]):

$$\pi(\inf I_\gamma) = \pi(\sup I_\gamma) = 1 - \gamma = 1 - P(I_\gamma).$$

This kind of transformation again yields a kind of cumulative distribution according to the ordering induced by the density  $p$ . Similar constructs can be found in the statistical literature (Birnbaum [74]). More recently Mauris et al. [76] noticed that starting from any family of nested sets around some characteristic point (the mean, the median, ...), the above equation yields a possibility measure dominating  $P$ . Well-known inequalities of probability theory, such as those of Chebyshev and Camp–Meidel, can also be viewed as possibilistic approximations of probability functions. It turns out that for symmetric uni-modal densities, each side of the optimal possibilistic transform is a convex function. Given such a probability density on a bounded interval  $[a, b]$ , the triangular fuzzy number whose core is the mode of  $p$  and the support is  $[a, b]$  is thus a possibility distribution dominating  $P$  regardless of its shape (and the tightest such distribution). These results justify the use of symmetric triangular fuzzy numbers as fuzzy counterparts to uniform probability distributions. They provide much tighter probability bounds than Chebyshev and Camp–Meidel inequalities for symmetric densities with bounded support. This setting is adapted to the modelling of sensor measurements [77]. These results are extended to more general distributions by Baudrit et al., [78], and provide a tool for representing poor probabilistic information.

### Subjective Possibility Distributions

The case of a subjective probability distribution is different. Indeed, the probability function is then supplied by an agent who is in some sense forced to express beliefs in this form due to rationality constraints, and the setting of exchangeable bets. However his actual knowledge may be

far from justifying the use of a single well-defined probability distribution. For instance in case of total ignorance about some value, apart from its belonging to an interval, the framework of exchangeable bets enforces a uniform probability distribution, on behalf of the principle of insufficient reason. Based on the setting of exchangeable bets, it is possible to define a subjectivist view of numerical possibility theory, that differs from the proposal of Walley [66]. The approach developed by Dubois Prade and Smets [79] relies on the assumption that when an agent constructs a probability measure by assigning prices to lotteries, this probability measure is actually induced by a belief function representing the agents actual state of knowledge. We assume that going from an underlying belief function to an elicited probability measure is achieved by means of the above mentioned pignistic transformation, changing focal sets into uniform probability distributions. The task is to reconstruct this underlying belief function under a minimal commitment assumption. In the paper [79], we pose and solve the problem of finding the least informative belief function having a given pignistic probability. We prove that it is unique and consonant, thus induced by a possibility distribution. This result exploits a simple partial ordering between belief functions comparing their information content, in agreement with the expected cardinality of random sets. The obtained possibility distribution can be defined as the converse of the pignistic transformation (which is one-to-one for possibility distributions). It is subjective in the same sense as in the subjectivist school in probability theory. However, it is the least biased representation of the agents state of knowledge compatible with the observed betting behavior. In particular it is less specific than the one constructed from the prediction intervals of an objective probability. This transformation was first proposed in [80] for objective probability, interpreting the empirical necessity of an event as summing the excess of probabilities of realizations of this event with respect to the probability of the most likely realization of the opposite event.

### Possibility Theory and Defuzzification

Possibilistic mean values can be defined using Choquet integrals with respect to possibility and necessity measures [65,81], and come close to defuzzification methods [82]. A fuzzy interval is a fuzzy set of reals whose membership function is unimodal and upper-semi continuous. Its  $\alpha$ -cuts are closed intervals. Interpreting a fuzzy interval  $M$ , associated to a possibility distribution  $\mu_M$ , as a family of probabilities, upper and lower mean values  $E^*(M)$



and  $E_*(M)$ , can be defined as [83]:

$$E_*(M) = \int_0^1 \inf M_\alpha d\alpha; \quad E^*(M) = \int_0^1 \sup M_\alpha d\alpha$$

where  $M_\alpha$  is the  $\alpha$ -cut of  $M$ .

Then the mean interval  $E(M) = [E_*(M), E^*(M)]$  of  $M$  is the interval containing the mean values of all random variables consistent with  $M$ , that is  $E(M) = \{E(P) | P \in \mathbf{P}(\mu_M)\}$ , where  $E(P)$  represents the expected value associated to the probability measure  $P$ . That the “mean value” of a fuzzy interval is an interval seems to be intuitively satisfactory. Particularly the mean interval of a (regular) interval  $[a, b]$  is this interval itself. The upper and lower mean values are linear with respect to the addition of fuzzy numbers. Define the addition  $M + N$  as the fuzzy interval whose cuts are  $M_\alpha + N_\alpha = \{s + t, s \in M_\alpha, t \in N_\alpha\}$  defined according to the rules of interval analysis. Then  $E(M + N) = E(M) + E(N)$ , and similarly for the scalar multiplication  $E(aM) = aE(M)$ , where  $aM$  has membership grades of the form  $\mu_M(s/a)$  for  $a \neq 0$ . In view of this property, it seems that the most natural defuzzification method is the middle point  $\hat{E}(M)$  of the mean interval (originally proposed by Yager [84]). Other defuzzification techniques do not generally possess this kind of linearity property.  $\hat{E}(M)$  has a natural interpretation in terms of simulation of a fuzzy variable [85], and is the mean value of the pignistic transformation of  $M$ . Indeed it is the mean value of the empirical probability distribution obtained by the random process defined by picking an element  $\alpha$  in the unit interval at random, and then an element  $s$  in the cut  $M_\alpha$  at random.

### Applications and Future Directions

Possibility theory has not been the main framework for engineering applications of fuzzy sets in the past. However, on the basis of its connections to symbolic artificial intelligence, to decision theory and to imprecise statistics, we consider that it has significant potential for further applied developments in a number of areas, including some where fuzzy sets are not yet always accepted. Only some directions are pointed out here.

1. Rules with exceptions can be modelled by means of conditional possibility [32], based on its capability to account for non-monotonic inference, as shown in Sect. “Non-monotonic Inference”. Possibility theory has also enabled a typology of fuzzy rules to be laid bare, distinguishing rules whose purpose is to propagate uncertainty through reasoning steps, from rules whose main purpose is similarity-based interpolation [86], depending on the choice of a many-valued implication
2. Possibility theory also offers a framework for preference modeling in constraint-directed reasoning. Both prioritized and soft constraints can be captured by possibility distributions expressing degrees of feasibility rather than plausibility [90]. Possibility offers a natural setting for fuzzy optimization whose aim is to balance the levels of satisfaction of multiple fuzzy constraints (instead of minimizing an overall cost) [91]. Qualitative decision criteria are particularly adapted to the handling of uncertainty in this setting. Applications of possibility theory-based decision-making can be found in scheduling [92,93,94,95].
3. Quantitative possibility theory is the natural setting for a reconciliation between probability and fuzzy sets. An important research direction is the comparison between fuzzy interval analysis [96] and random variable calculations with a view to unifying them [97]. Indeed, a current major concern, in for instance risk analysis studies, is to perform uncertainty propagation under poor data and without independence assumptions (see the papers in the special issue [98]). Finding the potential of possibilistic representations in computing conservative bounds for such probabilistic calculations is certainly a major challenge [99]. The active area of fuzzy random variables is also connected to this question [100].
4. One might also mention the well-known possibilistic clustering technique [101,102]. However, it is only loosely related to possibility theory. This name is due to the use of fuzzy clusters with (almost) unrestricted membership functions, that no longer form a usual fuzzy partition. But one might use it to generate genuine possibility distributions, where possibility derives from similarity, a point of view already mentioned before.

connective that models a rule. The bipolar view of information based on  $(\delta, \pi)$  pairs sheds new light on the debate between conjunctive and implicative representation of rules [87]. Representing a rule as a material implication focuses on counterexamples to rules, while using a conjunction between antecedent and consequent points out examples of the rule and highlights its positive content. Traditionally in fuzzy control and modelling, the latter representation is adopted, while the former is the logical tradition. Introducing fuzzy implicative rules in modelling accounts for constraints or landmark points the model should comply with (as opposed to observed data) [88]. The bipolar view of rules in terms of examples and counterexamples may turn out to be very useful when extracting fuzzy rules from data [89].



Other applications of possibility theory can be found in fields such as data analysis [103,104,105], database querying [106], diagnosis [107,108], belief revision [109], argumentation [110] case-based reasoning [111,112]. Lastly, possibility theory is also being studied from the point of view of its relevance in cognitive psychology. Experimental results [113] suggest that there are situations where people reason about uncertainty using the rules or possibility theory, rather than with those of probability theory.

## Bibliography

- Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst* 1:3–28
- Gaines BR, Kohout L (1975) Possible automata. In: *Proc Int Symp Multiple-Valued Logic*, Bloomington May 13–16. IEEE Press, pp 183–196
- Dubois D, Prade H (1998) Possibility theory: Qualitative and quantitative aspects. In: Gabbay DM, Smets PP (eds) *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol 1. Kluwer, Dordrecht, pp 169–226
- Dubois D, Nguyen HT, Prade H (2000) Fuzzy sets and probability: Misunderstandings, bridges and gaps. In: Dubois D, Prade H (eds) *Fundamentals of Fuzzy Sets*. Kluwer, Boston, pp 343–438
- Hughes GE, Cresswell MJ (1968) *An introduction to modal logic*. Methuen, London
- Shackle GLS (1961) *Decision, order and time in human affairs*, 2nd edn. Cambridge University Press, Cambridge
- Spohn W (1990) A general, nonprobabilistic theory of inductive reasoning. In: Shachter RD et al (eds) *Uncertainty in Artificial Intelligence*, vol 4. North Holland, Amsterdam, pp 149–158
- Lewis DL (1973) *Counterfactuals*. Basil Blackwell, Oxford
- Dubois D (1986) Belief structures, possibility theory and decomposable measures on finite sets. *Comput AI* 5:403–416
- Sudkamp T (2002) Similarity and the measurement of possibility. *Actes Rencontres Francophones sur la Logique Floue et ses Applications* (Montpellier, France). Cepadues Editions, Toulouse, pp 13–26
- Cohen LJ (1977) *The probable and the provable*. Clarendon, Oxford
- Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta MM, Ragade R, Yager RR (eds) *Advances in fuzzy set theory and applications*. North-Holland, Amsterdam, pp 3–18
- Zadeh LA (1982) Possibility theory and soft data analysis. In: Cobb L, Thrall R (eds) *Mathematical frontiers of social and policy sciences*. Westview Press, Boulder, pp 69–129
- Dubois D, Prade H (1980) *Fuzzy sets and systems: Theory and applications*. Academic Press, New York
- Dubois D, Prade H (1988) *Possibility theory*. Plenum, New York
- Klir GJ, Folger T (1988) *Fuzzy sets, uncertainty and information*. Prentice Hall, Englewood Cliffs
- Yager RR (1983) An introduction to applications of possibility theory. *Hum Syst Manag* 3:246–269
- Dubois D, Hajek P, Prade H (2000) Knowledge-driven versus data-driven logics. *J Log Lang Inf* 9:65–89
- Walley P (1996) Measures of uncertainty in expert systems. *Artif Intell* 83:1–58
- De Cooman G (1997) Possibility theory. Part I: Measure- and integral-theoretic groundwork; Part II: Conditional possibility; Part III: Possibilistic independence. *Int J Gen Syst* 25:291–371
- De Campos LM, Huete JF (1999) Independence concepts in possibility theory. *Fuzzy Sets Syst* 103:127–152 & 487–506
- Dubois DD, Farinas del Cerro L, Herzig A, Prade H (1997) Qualitative relevance and independence: A roadmap. In: *Proc of the 15th Inter Joint Conf on Artif Intell*, Nagoya, 23–29 August, 1997. Morgan Kaufmann, San Mateo, pp 62–67
- Borgelt C, Gebhardt J, Kruse R (2000) Possibilistic graphical models. In: Della Riccia G et al (eds) *Computational intelligence in data mining*. Springer, Wien, pp 51–68
- Benferhat S, Dubois D, Garcia L, Prade H (2002) On the transformation between possibilistic logic bases and possibilistic causal networks. *Int J Approx Reason* 29(2):135–173
- Ben Amor N, Benferhat S, Mellouli K (2003) Anytime propagation algorithm for min-based possibilistic graphs. *Soft Comput* 8(2):150–161
- Borgelt C, Kruse R (2003) Operations and evaluation measures for learning possibilistic graphical models. *Artif Intell* 148(1–2):385–418
- Gärdenfors P (1988) *Knowledge in flux*. MIT Press, Cambridge
- Dubois D, Prade H (1991) Epistemic entrenchment and possibilistic logic. *Artif Intell* 50:223–239
- Ben Amor N et al (2002) A theoretical framework for possibilistic independence in a weakly ordered setting. *Int J Uncert Fuzz & Knowl-B Syst* 10:117–155
- Lehmann D, Magidor M (1992) What does a conditional knowledge base entail? *Artif Intell* 55:1–60
- Dubois D, Fargier H, Prade H (2004) Ordinal and probabilistic representations of acceptance. *J Artif Intell Res* 22:23–56
- Benferhat S, Dubois D, Prade H (1997) Nonmonotonic reasoning, conditional objects and possibility theory *Artif Intell* 92:259–276
- Pearl J (1990) System Z: A natural ordering of defaults with tractable applications to default reasoning. In: *Proc 3rd Conf Theoretical Aspects of Reasoning About Knowledge*. Morgan Kaufmann, San Francisco, pp 121–135
- Dubois D, Lang J, Prade H (1994) Possibilistic logic. In: Gabbay DM et al (eds) *Handbook of logic in AI and logic programming*, vol 3. Oxford University Press, Oxford, pp 439–513
- Benferhat S, Dubois D, Prade H (1998) Practical handling of exception-tainted rules and independence information in possibilistic logic. *Appl Intell* 9:101–127
- Dubois D, Prade H (2004) Possibilistic logic: A retrospective and prospective view. *Fuzzy Sets Syst* 144:3–23
- Dubois D, Prade H (2007) Toward multiple-agent extensions of possibilistic logic. In: *Proc IEEE Inter Conf on Fuzzy Systems (FUZZ-IEEE (2007))*, London 23–26 July, 2007. pp 187–192
- Benferhat S, Dubois D, Garcia L, Prade H (2002) On the transformation between possibilistic logic bases and possibilistic causal networks. *Int J Approx Reason* 29:135–173
- Ben Amor N, Benferhat S (2005) Graphoid properties of qualitative possibilistic independence relations. *Int J Uncert Fuzz & Knowl-B Syst* 13:59–97
- Savage LJ (1972) *The foundations of statistics*. Dover, New York
- Yager RR (1979) Possibilistic decision making. *IEEE Trans Syst Man Cybern* 9:388–392

42. Whalen T (1984) Decision making under uncertainty with various assumptions about available information. *IEEE Trans Syst Man Cybern* 14:888–900
43. Grabisch M, Murofushi T, Sugeno M (eds) (2000) Fuzzy measures and integrals theory and applications. Physica, Heidelberg
44. Dubois D, Prade H, Sabbadin R (2000) Qualitative decision theory with Sugeno integrals. In: Grabisch M, Murofushi T, Sugeno M (eds) Fuzzy measures and integrals theory and applications. Physica, Heidelberg, pp 314–322
45. Dubois D, Prade H, Sabbadin R (2001) Decision-theoretic foundations of possibility theory. *Eur J Oper Res* 128:459–478
46. Dubois D, Fargier H, Perny P, Prade H (2003) Qualitative decision theory with preference relations and comparative uncertainty: An axiomatic approach. *Artif Intell* 148:219–260
47. Dubois D, Fargier H (2003) Qualitative decision rules under uncertainty. In: Della Riccia G et al (eds) Planning based on decision theory. CISM courses and Lectures, vol 472. Springer, Wien, pp 3–26
48. Fargier H, Sabbadin R (2005) Qualitative decision under uncertainty: Back to expected utility. *Artif Intell* 164:245–280
49. Dubois D (2006) Possibility theory and statistical reasoning. *Comput Stat Data Anal* 51(1):47–69
50. Dubois D, Prade H (1992) When upper probabilities are possibility measures. *Fuzzy Sets Syst* 49:s 65–74
51. Dubois D, Moral S, Prade H (1997) A semantics for possibility theory based on likelihoods. *J Math Anal Appl* 205:359–380
52. Shafer G (1987) Belief functions and possibility measures. In: Bezdek JC (ed) Analysis of fuzzy information, vol I: Mathematics and Logic. CRC Press, Boca Raton, pp 51–84
53. Benferhat S, Dubois D, Prade H (1999) Possibilistic and standard probabilistic semantics of conditional knowledge bases. *J Log Comput* 9:873–895
54. Maslov V (1987) Méthodes Opératoires. Mir Publications, Moscow
55. Kolokoltsov VN, Maslov VP (1997) Idempotent analysis and applications. Kluwer, Dordrecht
56. Puhalskii A (2001) Large deviations and idempotent probability. Chapman and Hall, Boca Raton
57. Nguyen HT, Bouchon-Meunier B (2003) Random sets and large deviations principle as a foundation for possibility measures. *Soft Comput* 8:61–70
58. De Cooman G, Aeyels D (1999) Supremum-preserving upper probabilities. *Inf Sci* 118:173–212
59. Walley P, De Cooman G (1999) A behavioural model for linguistic uncertainty. *Inf Sci* 134:1–37
60. Gebhardt J, Kruse R (1993) The context model. *Int J Approx Reason* 9:283–314
61. Joslyn C (1997) Measurement of possibilistic histograms from interval data. *Int J Gen Syst* 26:9–33
62. Neumaier A (2004) Clouds, fuzzy sets and probability intervals. *Reliab Comput* 10:249–272
63. De Baets B, Tsiporkova E, Mesiar R (1999) Conditioning in possibility with strict order norms. *Fuzzy Sets Syst* 106:221–229
64. Dubois D, Prade H (1997) Bayesian conditioning in possibility theory. *Fuzzy Sets Syst* 92:223–240
65. De Cooman G (2001) Integration and conditioning in numerical possibility theory. *Ann Math AI* 32:87–123
66. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, Boca Raton
67. Klir GJ (1990) A principle of uncertainty and information invariance. *Int J Gen Syst* 17:249–275
68. Geer JF, Klir GJ (1992) A mathematical analysis of information-preserving transformations between probabilistic and possibilistic formulations of uncertainty. *Int J Gen Syst* 20:143–176
69. Dubois D, Prade H, Sandri S (1993) On possibility/probability transformations. In: Lowen R, Roubens M (eds) Fuzzy logic: State of the art. Kluwer, Dordrecht, pp 103–112
70. Klir GJ, Parviz B (1992) Probability B-possibility transformations: A comparison. *Int J Gen Syst* 21:291–310
71. Dubois D, Prade H (1982) On several representations of an uncertain body of evidence. In: Gupta M, Sanchez E (eds) Fuzzy information and decision processes. North-Holland, Amsterdam, pp 167–181
72. Smets P (1990) Constructing the pignistic probability function in a context of uncertainty. In: Henrion M et al (eds) Uncertainty in artificial intelligence, vol 5. North-Holland, Amsterdam, pp 29–39
73. Dubois D, Prade H (2001) Smets new semantics for quantitative possibility theory. In: Proc ESQARU (2001), Toulouse, LNAI 2143. Springer, pp 410–421
74. Birnbaum ZW (1948) On random variables with comparable peakedness. *Ann Math Stat* 19:76–81
75. Dubois D, Huellermeier E (2005) A Notion of comparative probabilistic entropy based on the possibilistic specificity ordering. In: Godo L (ed) Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Proc of 8th European Conference, ECSQARU 2005, Barcelona, 6–8. Lecture Notes in Computer Science, vol 3571. Springer, Berlin
76. Dubois D, Foulloy L, Mauris G, Prade H (2004) Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliab Comput* 10:273–297
77. Mauris G, Lasserre V, Foulloy L (2000) Fuzzy modeling of measurement data acquired from physical sensors. *IEEE Trans Meas Instrum* 49:1201–1205
78. Baudrit C, Dubois D, Fargier H (2004) Practical representation of incomplete probabilistic information. In: Lopez-Diaz M et al (eds) Soft methods in probability and statistics. Proc 2nd Int Conf. Springer, Oviedo, pp 149–156
79. Dubois D, Prade H, Smets P (2003) A definition of subjective possibility. *Badania Operacyjne i Decyzje (Wroclaw)* 4:7–22
80. Dubois D, Prade H (1983) Unfair coins and necessity measures: A possibilistic interpretation of histograms. *Fuzzy Sets Syst* 10(1):15–20
81. Dubois D, Prade H (1985) Evidence measures based on fuzzy information. *Automatica* 21:547–562
82. Van Leekwijck W, Kerre EE (2001) Defuzzification: Criteria and classification. *Fuzzy Sets Syst* 108:303–314
83. Dubois D, Prade H (1987) The mean value of a fuzzy number. *Fuzzy Sets Syst* 24:279–300
84. Yager RR (1981) A procedure for ordering fuzzy subsets of the unit interval. *Inf Sci* 24:143–161
85. Chanas S, Nowakowski M (1988) Single value simulation of fuzzy variable. *Fuzzy Sets Syst* 25:43–57
86. Dubois D, Prade H (1996) What are fuzzy rules and how to use them. *Fuzzy Sets Syst* 84:169–185
87. Dubois D, Prade H, Ughetto L (2003) A new perspective on reasoning with fuzzy rules. *Int J Intell Syst* 18:541–567
88. Galichet S, Dubois D, Prade H (2004) Imprecise specification of ill-known functions using gradual rules. *Int J Approx Reason* 35:205–222

89. Dubois D, Huellermeier E, Prade H (2003) A note on quality measures for fuzzy association rules. In: De Baets B, Bilgic T (eds) *Fuzzy sets and systems*. Proc of the 10th Int Fuzzy Systems Assoc World Congress IFSA, Istanbul, 2003, LNAI 2715. Springer, pp 346–353
90. Dubois D, Fargier H, Prade H (1996) Possibility theory in constraint satisfaction problems: Handling priority, preference and uncertainty. *Appl Intell* 6:287–309
91. Dubois D, Fortemps P (1999) Computing improved optimal solutions to max-min flexible constraint satisfaction problems. *Eur J Oper Res* 118:95–126
92. Dubois D, Fargier H, Prade H (1995) Fuzzy constraints in job-shop scheduling. *J Intell Manuf* 6:215–234
93. Slowinski R, Hapke M (eds) (2000) *Scheduling under fuzziness*. Physica, Heidelberg
94. Chanas S, Zielinski P (2001) Critical path analysis in the network with fuzzy activity times. *Fuzzy Sets Syst* 122:195–204
95. Chanas S, Dubois D, Zielinski P (2002) Necessary criticality in the network with imprecise activity times. *IEEE Trans Man Mach Cybern* 32:393–407
96. Dubois D, Kerre E, Mesiar R, Prade H (2000) Fuzzy interval analysis. In: Dubois D, Prade H (eds) *Fundamentals of fuzzy sets*. Kluwer, Boston, pp 483–581
97. Dubois D, Prade H (1991) Random sets and fuzzy interval analysis. *Fuzzy Sets Syst* 42:87–101
98. Helton JC, Oberkampf WL (eds) (2004) *Alternative Representations of Epistemic Uncertainty*. Reliability Engineering and Systems Safety, vol 85. Elsevier, Amsterdam, p 369
99. Guyonnet D et al (2003) Hybrid approach for addressing uncertainty in risk assessments. *J Env Eng* 129:68–78
100. Gil M (ed) (2001) *Fuzzy random variables*. Inf Sci 133
101. Krishnapuram R, Keller J (1993) A possibilistic approach to clustering. *IEEE Trans Fuzzy Syst* 1:98–110
102. Bezdek J, Keller J, Krishnapuram R, Pal N (1999) Fuzzy models and algorithms for pattern recognition and image processing. In: *The Handbooks of Fuzzy Sets Series*. Kluwer, Boston
103. Wolkenhauer O (1998) Possibility theory with applications to data analysis. Research Studies Press, Chichester
104. Tanaka H, Guo PJ (1999) Possibilistic data analysis for operations research. Physica, Heidelberg
105. Borgelt C, Gebhardt J, Kruse R (2000) Possibilistic graphical models. In: Della Riccia G et al (eds) *Computational intelligence in data mining*. CISM Series, vol N408. Springer, Berlin
106. Bosc P, Prade H (1997) An introduction to the fuzzy set and possibility theory-based treatment of soft queries and uncertain of imprecise databases. In: Smets P, Motro A (eds) *Uncertainty management in information systems*. Kluwer, Dordrecht, pp 285–324
107. Cayrac D, Dubois D, Prade H (1996) Handling uncertainty with possibility theory and fuzzy sets in a satellite fault diagnosis application. *IEEE Trans Fuzzy Syst* 4:251–269
108. Boverie S et al (2002) Online diagnosis of engine dyno test benches: A possibilistic approach. Proc 15th Eur Conf on Artificial Intelligence, Lyon. IOS Press, Amsterdam, p 658–662
109. Benferhat S, Dubois D, Prade H, Williams M-A (2002) A practical approach to revising prioritized knowledge bases. *Stud Log* 70:105–130
110. Amgoud L, Prade H (2004) Reaching agreement through argumentation: A possibilistic approach. In: Proc of the 9th Int Conf on Principles of Knowledge Representation and Reasoning (KR'04), Whistler. AAAI Press, Cambridge, pp 175–182
111. Dubois D, Huellermeier E, Prade H (2002) Fuzzy set-based methods in instance-based reasoning. *IEEE Trans Fuzzy Syst* 10:322–332
112. Huellermeier E, Dubois D, Prade H (2002) Model adaptation in possibilistic instance-based reasoning. *IEEE Trans Fuzzy Syst* 10:333–339
113. Raufaste E, Da Silva R, Neves C (2003) Mariné testing the descriptive validity of possibility theory in human judgements of uncertainty. *Artif Intell* 148:197–218

## Pressure and Equilibrium States in Ergodic Theory

JEAN-RENÉ CHAZOTTES<sup>1</sup>, GERHARD KELLER<sup>2</sup>

<sup>1</sup> Centre de Physique Théorique,  
CNRS/École Polytechnique,  
Palaiseau, France

<sup>2</sup> Department Mathematik, Universität  
Erlangen-Nürnberg, Erlangen, Germany

### Article Outline

Glossary

Definition of the Subject

Introduction

Warming Up: Thermodynamic Formalism  
for Finite Systems

Shift Spaces, Invariant Measures and Entropy

The Variational Principle: A Global Characterization  
of Equilibrium

The Gibbs Property: A Local Characterization  
of Equilibrium

Examples on Shift Spaces

Examples from Differentiable Dynamics

Nonequilibrium Steady States and Entropy Production

Some Ongoing Developments and Future Directions

Bibliography

### Glossary

**Dynamical system** In this article: a continuous transformation  $T$  of a compact metric space  $X$ . For each  $x \in X$ , the transformation  $T$  generates a trajectory  $(x, Tx, T^2x, \dots)$ .

**Invariant measure** In this article: a probability measure  $\mu$  on  $X$  which is invariant under the transformation  $T$ , i. e., for which  $\langle f \circ T, \mu \rangle = \langle f, \mu \rangle$  for each continuous  $f: X \rightarrow \mathbb{R}$ . Here  $\langle f, \mu \rangle$  is a short-hand notation for  $\int_X f d\mu$ . The triple  $(X, T, \mu)$  is called a measure-preserving dynamical system.

**Ergodic theory** Ergodic theory is the mathematical theory of measure-preserving dynamical systems.

**Entropy** In this article: the maximal rate of information gain per time that can be achieved by coarse-grained observations on a measure-preserving dynamical system. This quantity is often denoted  $h(\mu)$ .

**Equilibrium state** In general, a given dynamical system  $T: X \rightarrow X$  admits a huge number of invariant measures. Given some continuous  $\phi: X \rightarrow \mathbb{R}$  (“potential”), those invariant measures which maximize a functional of the form  $F(\mu) = h(\mu) + \langle \phi, \mu \rangle$  are called “equilibrium states” for  $\phi$ .

**Pressure** The maximum of the functional  $F(\mu)$  is denoted by  $P(\phi)$  and called the “topological pressure” of  $\phi$ , or simply the “pressure” of  $\phi$ .

**Gibbs state** In many cases, equilibrium states have a local structure that is determined by the local properties of the potential  $\phi$ . They are called “Gibbs states”.

**Sinai–Ruelle–Bowen measure** Special equilibrium or Gibbs states that describe the statistics of the attractor of certain smooth dynamical systems.

## Definition of the Subject

Gibbs and equilibrium states of one-dimensional lattice models in statistical physics play a prominent role in the statistical theory of chaotic dynamics. They first appear in the ergodic theory of certain differentiable dynamical systems, called “uniformly hyperbolic systems”, mainly Anosov and Axiom A diffeomorphisms (and flows). The central idea is to “code” the orbits of these systems into (infinite) symbolic sequences of symbols by following their history on a finite partition of their phase space. This defines a nice shift dynamical system called a subshift of finite type or a topological Markov chain. Then the construction of their “natural” invariant measures and the study of their properties are carried out at the symbolic level by constructing certain equilibrium states in the sense of statistical mechanics which turn out to be also Gibbs states. The study of uniformly hyperbolic systems brought out several ideas and techniques which turned out to be extremely fruitful for the study of more general systems. Let us mention the concept of Markov partition and its avatars, the very important notion of SRB measure (after Sinai, Ruelle, and Bowen) and transfer operators. Recently, there was a revival of interest in Axiom A systems as models to understand nonequilibrium statistical mechanics.

## Introduction

Our goal is to present the basic results on one-dimensional Gibbs and equilibrium states viewed as special invariant measures on symbolic dynamical systems, and then to describe without technicalities a sample of results they allow

to obtain for certain differentiable dynamical systems. We hope that this contribution will illustrate the symbiotic relationship between ergodic theory and statistical mechanics, and also information theory.

We start by putting Gibbs and equilibrium states in a general perspective. The theory of Gibbs states and equilibrium states, or Thermodynamic Formalism, is a branch of rigorous Statistical Physics. The notion of a Gibbs state dates back to R.L. Dobrushin (1968–1969) [17,18,19,20] and O.E. Lanford and D. Ruelle (1969) [41] who proposed it as a mathematical idealization of an equilibrium state of a physical system which consists of a very large number of interacting components. For a finite number of components, the foundations of statistical mechanics were already laid in the nineteenth century. There was the well-known Maxwell–Boltzmann–Gibbs formula for the equilibrium distribution of a physical system with given energy function. From the mathematical point of view, the intrinsic properties of very large objects can be made manifest by performing suitable limiting procedures. Indeed, the crucial step made in the 1960s was to define the notion of a Gibbs measure or Gibbs state for a system with an infinite number of interacting components. This was done by the familiar probabilistic idea of specifying the interdependence structure by means of a suitable class of conditional probabilities built up according to the Maxwell–Boltzmann–Gibbs formula [29]. Notice that Gibbs states are often called “DLR states” in honor of Dobrushin, Lanford, and Ruelle. The remarkable aspect of this construction is the fact that a Gibbs state for a given type of interaction may fail to be unique. In physical terms, this means that a system with this interaction can take several distinct equilibria. The phenomenon of nonuniqueness of a Gibbs measure can thus be interpreted as a phase transition. Therefore, the conditions under which an interaction leads to a unique or to several Gibbs measures turns out to be of central importance. While Gibbs states are defined locally by specifying certain conditional probabilities, equilibrium states are defined globally by a *variational principle*: they maximize the entropy of the system under the (linear) constraint that the mean energy is fixed. Gibbs states are always equilibrium states, but the two notions do not coincide in general. However, for a class of sufficiently regular interactions, equilibrium states are also Gibbs states.

In the effort of trying to understand phase transitions, simplified mathematical models were proposed, the most famous one being undoubtedly the Ising model. This is an example of a lattice model. The set of configurations of a lattice model is  $X := A^{\mathbb{Z}^d}$ , where  $A$  is a finite set, which is invariant by “spatial” translations. For the physical in-



terpretation,  $X$  can be thought, for instance, as the set of infinite configurations of a system of spins on a crystal lattice  $\mathbb{Z}^d$  and one may take  $A = \{-1, +1\}$ , i. e., spins can take two orientations, “up” and “down”. The Ising model is defined by specifying an interaction (or potential) between spins and studying the corresponding (translation-invariant) Gibbs states. The striking phenomenon is that for  $d = 1$  there is a unique Gibbs state (in fact a Markov measure) whereas if  $d \geq 2$ , there may be several Gibbs states although the interaction is very simple [29].

Equilibrium states and Gibbs states of one-dimensional lattice models ( $d = 1$ ) played a prominent role in understanding the ergodic properties of certain types of differentiable dynamical systems, namely uniformly hyperbolic systems, Axiom A diffeomorphisms in particular. The link between one-dimensional lattice systems and dynamical systems is made by *symbolic dynamics*. Informally, symbolic dynamics consists of replacing the orbits of the original system by its history on a finite partition of its phase space labeled by the elements of the “alphabet”  $A$ . Therefore, each orbit of the original system is replaced by an infinite sequence of symbols, i. e., by an element of the set  $A^{\mathbb{Z}}$  or  $A^{\mathbb{N}}$ , depending on whether the map describing the dynamics is invertible or not. The action of the map on an initial condition is then easily seen to correspond to the translation (or shift) of its associated symbolic sequence. In general there is no reason to get all sequences of  $A^{\mathbb{Z}}$  or  $A^{\mathbb{N}}$ . Instead one gets a closed invariant subset  $X$  (a subshift) which can be very complicated. For a certain class of dynamical systems the partition can be successfully chosen so as to form a *Markov partition*. In this case, the dynamical system under consideration can be coded by a *subshift of finite type* (also called a *topological Markov chain*) which is a very nice symbolic dynamical system. Then one can play the game of statistical physics: for a given continuous, real-valued function (a “potential”) on  $X$ , construct the corresponding Gibbs states and equilibrium states. If the potential is regular enough, one expects uniqueness of the Gibbs state and that it is also the unique equilibrium state for this potential. This circle of ideas – ranging from Gibbs states on finite systems over invariant measures on symbolic systems and their (Shannon-)entropy with a digression to Kolmogorov–Chaitin complexity to equilibrium states and Gibbs states on subshifts of finite type – is presented in the next four sections.

At this point it should be remembered that the objects which can actually be observed are not equilibrium states (they are measures on  $X$ ) but individual symbol sequences in  $X$ , which reflect more or less the statistical properties of an equilibrium state. Indeed, most sequences reflect these properties very well, but there are also rare sequences that

look quite different. Their properties are described by *large deviations principles* which are not discussed in the present article. We shall indicate some references along the way.

In Sects. “[Examples on Shift Spaces](#)” and “[Examples from Differentiable Dynamics](#)” we present a selection of important examples: measure of maximal entropy, Markov measures and Hofbauer’s example of nonuniqueness of equilibrium state; uniformly expanding Markov maps of the interval, interval maps with an indifferent fixed point, Anosov diffeomorphisms and Axiom A attractors with Sinai–Ruelle–Bowen measures, and Bowen’s formula for the Hausdorff dimension of conformal repellers. As we shall see, Sinai–Ruelle–Bowen measures are the only physically observable measures and they appear naturally in the context of nonuniformly hyperbolic diffeomorphisms [71].

A revival of the interest to Anosov and Axiom A systems occurred in statistical mechanics in the 1990s. Several physical phenomena of nonequilibrium origin, like entropy production and chaotic scattering, were modeled with the help of those systems (by G. Gallavotti, P. Gaspard, D. Ruelle, and others). This new interest led to new results about old Anosov and Axiom A systems, see, e. g., [15] for a survey and references. In Sect. “[Nonequilibrium Steady States and Entropy Production](#)”, we give a very brief account of *entropy production* in the context of Anosov systems which highlights the role of *relative entropy*.

This article is a little introduction to a vast subject in which we have tried to put forward some aspects not previously described in other expository texts. For readers willing to deepen their understanding of equilibrium and Gibbs states, there are the classic monographs by Bowen [6] and by Ruelle [58], the monograph by one of us [38], and the survey article by Chernov [15] (where Anosov and Axiom A flows are reviewed). Those texts are really complementary.

## Warming Up: Thermodynamic Formalism for Finite Systems

We introduce the thermodynamic formalism in an elementary context, following Jaynes [34]. In this view, entropy, in the sense of information theory, is the central concept.

Incomplete knowledge about a system is conveniently described in terms of probability distributions on the set of its possible states. This is particularly simple if the set of states, call it  $X$ , is finite. Then the equidistribution on  $X$  describes complete lack of knowledge, whereas a probability vector that assigns probability 1 to one single state and



probability 0 to all others represents maximal information about the system. A well-established measure of the amount of uncertainty represented by a probability distribution  $\nu = (\nu(x))_{x \in X}$  is its *entropy*

$$H(\nu) := - \sum_{x \in X} \nu(x) \log \nu(x),$$

which is zero if the probability is concentrated in one state and which attains its maximum value  $\log |X|$  if  $\nu$  is the equidistribution on  $X$ , i. e., if  $\nu(x) = |X|^{-1}$  for all  $x \in X$ . In this completely elementary context we will explore two concepts whose generalizations are central to the theory of equilibrium states in ergodic theory:

- Equilibrium distributions – defined in terms of a variational problem.
- The Gibbs property of equilibrium distributions.

The only mathematical prerequisite for this section are calculus and some elements from probability theory.

### Equilibrium Distributions and the Gibbs Property

Suppose that a finite system can be observed through a function  $U: X \rightarrow \mathbb{R}$  (an “observable”), and that we are looking for a probability distribution  $\mu$  which maximizes entropy among all distributions  $\nu$  with a prescribed expected value  $\langle U, \nu \rangle := \sum_{x \in X} \nu(x) U(x)$  for the observable  $U$ . This means we have to solve a variational problem under constraints:

$$H(\mu) = \max\{H(\nu) : \langle U, \nu \rangle = E\}. \quad (1)$$

As the function  $\nu \mapsto H(\nu)$  is strictly concave, there is a unique maximizing probability distribution  $\mu$  provided the value  $E$  can be attained at all by some  $\langle U, \nu \rangle$ . In order to derive an explicit formula for this  $\mu$  we introduce a Lagrange multiplier  $\beta \in \mathbb{R}$  and study, for each  $\beta$ , the unconstrained problem

$$H(\mu_\beta) + \langle \beta U, \mu_\beta \rangle = p(\beta U) := \max_{\nu} (H(\nu) + \langle \beta U, \nu \rangle). \quad (2)$$

In analogy to the convention in ergodic theory we call  $p(\beta U)$  the *pressure* of  $\beta U$  and the maximizer  $\mu_\beta$  the corresponding *equilibrium distribution* (synonymously *equilibrium state*).

The equilibrium distribution  $\mu_\beta$  satisfies

$$\mu_\beta(x) = \exp(-p(\beta U) + \beta U(x)) \quad \text{for all } x \in X \quad (3)$$

as an elementary calculation using Jensen’s inequality for the strictly convex function  $t \mapsto -t \log t$  shows:

$$\begin{aligned} H(\nu) + \langle \beta U, \nu \rangle &= \sum_{x \in X} \nu(x) \log \frac{e^{\beta U(x)}}{\nu(x)} \\ &\leq \log \sum_{x \in X} \nu(x) \frac{e^{\beta U(x)}}{\nu(x)} \\ &= \log \sum_{x \in X} e^{\beta U(x)}, \end{aligned}$$

with equality if and only if  $e^{\beta U}$  is a constant multiple of  $\nu$ . The observation that  $\nu = \mu_\beta$  is a maximizer proves at the same time that  $p(\beta U) = \log \sum_{x \in X} e^{\beta U(x)}$ .

The equality expressed in (3) is called the *Gibbs property* of  $\mu_\beta$ , and we say that  $\mu_\beta$  is a Gibbs distribution if we want to stress this property.

In order to solve the constrained problem (1) it remains to show that there is a unique multiplier  $\beta = \beta(E)$  such that  $\langle U, \mu_\beta \rangle = E$ . This follows from the fact that the map  $\beta \mapsto \langle U, \mu_\beta \rangle$  maps the real line monotonically onto the interval  $(\min U, \max U)$  which, in turn, is a direct consequence of the formulas for the first and second derivative of  $p(\beta U)$  w.r.t  $\beta$ :

$$\frac{dp}{d\beta} = \langle U, \mu_\beta \rangle, \quad \frac{d^2 p}{d\beta^2} = \langle U^2, \mu_\beta \rangle - \langle U, \mu_\beta \rangle^2. \quad (4)$$

As the second derivative is nothing but the variance of  $U$  under  $\mu_\beta$ , it is strictly positive (except when  $U$  is a constant function), so that  $\beta \mapsto \langle U, \mu_\beta \rangle$  is indeed strictly increasing. Observe also that  $dp/d\beta$  is indeed the directional derivative of  $p: \mathbb{R}^{|A|} \rightarrow \mathbb{R}$  in direction  $U$ . Hence the first identity in (4) can be rephrased as:  $\mu_\beta$  is the gradient at  $\beta U$  of the function  $p$ .

A similar analysis can be performed for an  $\mathbb{R}^d$ -valued observable  $U$ . In that case a vector  $\beta \in \mathbb{R}^d$  of Lagrange multipliers is needed to satisfy the  $d$  linear constraints.

### Systems on a Finite Lattice

We now assume that the system has a lattice structure, modeling its extension in space, for instance. The system can be in different states at different positions. More specifically, let  $\mathbb{L}_n = \{0, 1, \dots, n-1\}$  be a set of  $n$  positions in space, let  $A$  be a finite set of states that can be attained by the system at each of its sites, and denote by  $X := A^{\mathbb{L}_n}$  the set of all configurations of states from  $A$  at positions of  $\mathbb{L}_n$ . It is helpful to think of  $X$  as the set of all words of length  $n$  over the alphabet  $A$ . We focus on observables  $U_n$  which are sums of many local contributions in the sense that  $U_n(a_0 \dots a_{n-1}) = \sum_{i=0}^{n-1} \phi(a_i \dots a_{i+r-1})$  for

some “local observable”  $\phi: A^r \rightarrow \mathbb{R}$ . (The index  $i + r - 1$  has to be taken modulo  $n$ .) In terms of  $\phi$  the maximizing measure can be written as

$$\mu_\beta(a_0 \dots a_{n-1}) = \exp \left( -nP(\beta\phi) + \beta \sum_{i=0}^{n-1} \phi(a_i \dots a_{i+r-1}) \right), \quad (5)$$

where  $P(\beta\phi) := n^{-1}p(\beta U_n)$ . A first immediate consequence of (5) is the invariance of  $\mu_\beta$  under a cyclic shift of its argument, namely  $\mu_\beta(a_1 \dots a_{n-1}a_0) = \mu_\beta(a_0 \dots a_{n-1})$ . Therefore, we can restrict the maximizations in (1) and (2) to probability distributions  $\nu$  which are invariant under cyclic translations which yields

$$\begin{aligned} P(\beta\phi) &= \max_{\nu} (n^{-1}H(\nu) + \langle \beta\phi, \nu \rangle) \\ &= n^{-1}H(\mu_\beta) + \langle \beta\phi, \mu_\beta \rangle. \end{aligned} \quad (6)$$

If the local observable  $\phi$  depends only on one coordinate,  $\mu_\beta$  turns out to be a product measure:

$$\mu_\beta(a_0 \dots a_{n-1}) = \prod_{i=0}^{n-1} \exp(-P(\beta\phi) + \beta\phi(a_i)).$$

Indeed, comparison with (3) shows that  $\mu_\beta$  is the  $n$ -fold product of the probability distribution  $\mu_\beta^{\text{loc}}$  on  $A$  that maximizes  $H(\nu) + \beta\nu(\phi)$  among all distributions  $\nu$  on  $A$ . It follows that  $n^{-1}H(\mu_\beta) = H(\mu_\beta^{\text{loc}})$  so that (6) implies  $P(\beta\phi) = p(\beta\phi)$  for observables  $\phi$  that depend only on one coordinate.

### Shift Spaces, Invariant Measures and Entropy

We now turn to *shift dynamical systems* over a finite alphabet  $A$ .

#### Symbolic Dynamics

We start by fixing some notation. Let  $\mathbb{N}$  denote the set  $\{0, 1, 2, \dots\}$ . In the sequel we need

- a finite set  $A$  (the “alphabet”),
- the set  $A^{\mathbb{N}}$  of all infinite sequences over  $A$ , i. e., the set of all  $\underline{x} = x_0x_1\dots$  with  $x_n \in A$  for all  $n \in \mathbb{N}$ ,
- the translation (or shift)  $\sigma: A^{\mathbb{N}} \rightarrow A^{\mathbb{N}}$ ,  $(\sigma\underline{x})_n = x_{n+1}$ , for all  $n \in \mathbb{N}$ ,
- a shift invariant subset  $X = \sigma(X)$  of  $A^{\mathbb{N}}$ . With a slight abuse of notation we denote the restriction of  $\sigma$  to  $X$  by  $\sigma$  again.

We mention two interpretations of the dynamics of  $\sigma$ : it can describe the evolution of a system with state space  $X$

in discrete time steps (this is the prevalent interpretation if  $\sigma: X \rightarrow X$  is obtained as a symbolic representation of another dynamical system), or it can be the spatial translation of the configuration of a system on an infinite lattice (generalizing the point of view from Subsect. “[Systems on a Finite Lattice](#)” above). In the latter case one usually looks at the shift on the two-sided shift space  $A^{\mathbb{Z}}$ , for which the theory is nearly identical.

On  $A^{\mathbb{N}}$  one can define a metric  $d$  by

$$\begin{aligned} d(\underline{x}, \underline{y}) &:= 2^{-N(\underline{x}, \underline{y})} \\ \text{where } N(\underline{x}, \underline{y}) &:= \min\{k \in \mathbb{N} : x_k \neq y_k\}. \end{aligned} \quad (7)$$

Hence  $d(\underline{x}, \underline{y}) = 1$  if and only if  $x_0 \neq y_0$ , and  $d(\underline{x}, \underline{x}) = 0$  upon agreeing that  $N(\underline{x}, \underline{x}) = \infty$  and  $2^{-\infty} = 0$ . Equipped with this metric,  $A^{\mathbb{N}}$  becomes a compact metric space and  $\sigma$  is easily seen to be a continuous surjection of  $A^{\mathbb{N}}$ . Finally, if  $X$  is a closed subset of  $A^{\mathbb{N}}$ , we call the restriction  $\sigma: X \rightarrow X$ , which is again a continuous surjection, a shift dynamical system. We remark that  $d$  generates on  $A^{\mathbb{N}}$  the product topology of the discrete topology on  $A$ , just as many variants of  $d$  do. For more details ▶ [Symbolic Dynamics](#). As usual,  $C(X)$  denotes the space of real-valued continuous functions on  $X$  equipped with the supremum norm  $\|\cdot\|_\infty$ .

#### Invariant Measures

A probability distribution  $\nu$  (or simply distribution) on  $X$  is a Borel probability measure on  $X$ . It is unambiguously specified by its values  $\nu[a_0 \dots a_{n-1}]$  ( $n \in \mathbb{N}$ ,  $a_i \in A$ ) on *cylinder sets*

$$\begin{aligned} [a_0 \dots a_{n-1}] &:= \{\underline{x} \in X : x_i = a_i \text{ for all } i = 0, \dots, n-1\}. \end{aligned}$$

Any bounded and measurable  $f: X \rightarrow \mathbb{R}$  (in particular any  $f \in C(X)$ ) can be integrated by any distribution  $\nu$ . To stress the linearity of the integral in both, the integrand and the integrator, we use the notation

$$\langle f, \nu \rangle := \int_X f d\nu.$$

In probabilistic terms,  $\langle f, \nu \rangle$  is the expectation of the observable  $f$  under  $\nu$ . The set  $\mathcal{M}(X)$  of all probability distributions is compact in the weak topology, the coarsest topology on  $\mathcal{M}(X)$  for which  $\nu \mapsto \langle f, \nu \rangle$  is continuous for all  $f \in C(X)$ , ▶ [Measure Preserving Systems](#), Subsect. “Existence of Invariant Measures”. (Note that in functional analysis this is called the weak-\* topology.) Henceforth we will use both terms, “measure” and “distribution”, if we talk about probability distributions.

A measure  $\nu$  on  $X$  is *invariant* if expectations of observables are unchanged under the shift, i. e., if

$$\langle f \circ \sigma, \nu \rangle = \langle f, \nu \rangle$$

for all bounded measurable  $f: X \rightarrow \mathbb{R}$ .

The set of all invariant measures is denoted by  $\mathcal{M}_\sigma(X)$ . As a closed subset of  $\mathcal{M}(X)$  it is compact in the weak topology. Of special importance among all invariant measures  $\nu$  are the ergodic ones which can be characterized by the property that, for all bounded measurable  $f: X \rightarrow \mathbb{R}$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(\sigma^k \underline{x}) = \langle f, \nu \rangle$$

for  $\nu$ -a. e. (almost every)  $\underline{x}$ , (8)

i. e., for a set of  $\underline{x}$  of  $\nu$ -measure one. They are the indecomposable “building blocks” of all other measures in  $\mathcal{M}_\sigma(X)$ , [► Measure Preserving Systems](#) or [► Ergodic Theorems](#). The almost everywhere convergence in (8) is Birkhoff’s ergodic Theorem [► Ergodic Theorems](#), the constant limit characterizes the ergodicity of  $\nu$ .

### Entropy of Invariant Measures

We give a brief account of the definition and basic properties of the entropy of an invariant measure  $\nu$ . For details and the generalization of this concept to general dynamical systems we refer to [► Entropy in Ergodic Theory](#) or [37], and to [36] for an historical account.

Let  $\nu \in \mathcal{M}_\sigma(X)$ . For each  $n > 0$  the cylinder probabilities  $\nu[a_0 \dots a_{n-1}]$  give rise to a probability distribution on the finite set  $A^{\mathbb{L}_n}$ , see Sect. “[Warming Up: Thermodynamic Formalism for Finite Systems](#)”, so

$$H_n(\nu) := - \sum_{a_0, \dots, a_{n-1} \in A} \nu[a_0 \dots a_{n-1}] \log \nu[a_0 \dots a_{n-1}]$$

is well defined. Invariance of  $\nu$  guarantees that the sequence  $(H_n(\nu))_{n>0}$  is *subadditive*, i. e.,  $H_{k+n}(\nu) \leq H_k(\nu) + H_n(\nu)$ , and an elementary argument shows that the limit

$$h(\nu) := \lim_{n \rightarrow \infty} \frac{1}{n} H_n(\nu) \in [0, \log |A|] \quad (9)$$

exists and equals the infimum of the sequence. We simply call it the *entropy* of  $\nu$ . (Note that for general subshifts  $X$  many of the cylinder sets  $[a_0 \dots a_{n-1}] \subseteq X$  are empty. But, because of the continuity of the function  $t \mapsto t \log t$  at  $t = 0$ , we may set  $0 \log 0 = 0$ , and, hence, this does not affect the definition of  $H_n(\nu)$ .)

The entropy  $h(\nu)$  of an ergodic measure  $\nu$  can be observed along a “typical” trajectory. That is the content of the following theorem, sometimes called the “ergodic theorem of information theory” [► Entropy in Ergodic Theory](#).

### Theorem (Shannon–McMillan–Breiman Theorem)

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \nu[x_0 \dots x_{n-1}] = -h(\nu) \quad \text{for } \nu\text{-a. e. } \underline{x}. \quad (10)$$

Observe that (9) is just the integrated version of this statement. A slightly weaker reformulation of this theorem (again for ergodic  $\nu$ ) is known as the “asymptotic equipartition property”.

### Asymptotic Equipartition Property

Given (arbitrarily small)  $\epsilon > 0$  and  $\alpha > 0$ , one can, for each sufficiently large  $n$ , partition the set  $A^n$  into a set  $\mathcal{T}_n$  of typical words and a set  $\mathcal{E}_n$  of exceptional words such that each  $a_0 \dots a_{n-1} \in \mathcal{T}_n$  satisfies

$$e^{-n(h(\nu)+\alpha)} \leq \nu[a_0 \dots a_{n-1}] \leq e^{-n(h(\nu)-\alpha)} \quad (11)$$

and the total probability  $\sum_{a_0 \dots a_{n-1} \in \mathcal{E}_n} \nu[a_0 \dots a_{n-1}]$  of the exceptional words is at most  $\epsilon$ .

### A Short Digression on Complexity

Kolmogorov [40] and Chaitin [14] introduced the concept of complexity of an infinite sequence of symbols. Very roughly it is defined as follows: First, the complexity  $K(x_0 \dots x_{n-1})$  of a finite word in  $A^n$  is defined as the bit length of the shortest program that causes a suitable general purpose computer (say a PC or, for the mathematically minded reader, a Turing machine) to print out this word. Then the complexity of an infinite sequence is defined as  $K(\underline{x}) := \limsup_{n \rightarrow \infty} \frac{1}{n} K(x_0 \dots x_{n-1})$ . Of course, the definition of  $K(x_0 \dots x_{n-1})$  depends on the particular computer, but as any two general purpose computers can be programmed to simulate each other (by some finite piece of software), the limit  $K(\underline{x})$  is machine independent. It is the optimal compression factor for long initial pieces of a sequence  $\underline{x}$  that still allows complete reconstruction of  $\underline{x}$  by an algorithm. Brudno [8] showed:

If  $X \subseteq A^{\mathbb{N}}$  and  $\nu \in \mathcal{M}_\sigma(X)$  is ergodic, then  $K(\underline{x}) = \frac{1}{\log 2} h(\nu)$  for  $\nu$ -a. e.  $\underline{x} \in X$ .

### Entropy as a Function of the Measure

An important technical remark for the further development of the theory is that the entropy function

$h: \mathcal{M}_\sigma(X) \rightarrow [0, \infty)$  is *upper semicontinuous*. This means that all sets  $\{\nu: h(\nu) \geq t\}$  with  $t \in \mathbb{R}$  are closed and hence compact. In particular, upper semicontinuous functions attain their supremum. Indeed, suppose a sequence  $\nu_k \in \mathcal{M}_\sigma(X)$  converges weakly to some  $\nu \in \mathcal{M}_\sigma(X)$  and  $h(\nu_k) \geq t$  for all  $k$  so that also  $\frac{1}{n}H_n(\nu_k) \geq t$  for all  $n$  and  $k$ . As  $H_n(\nu)$  is an expression that depends continuously on the probabilities of the finitely many cylinders  $[a_0 \dots a_{n-1}]$  and as the indicator functions of these sets are continuous,  $\frac{1}{n}H_n(\nu) = \lim_{k \rightarrow \infty} \frac{1}{n}H_n(\nu_k) \geq t$ , hence  $h(\nu) \geq t$  in the limit  $n \rightarrow \infty$ .

A word of caution seems in order: the entropy function is rarely continuous. For example, on the full shift  $X = A^\mathbb{N}$  each invariant measure, whatever its entropy is, can be approximated in the weak topology by equidistributions on periodic orbits. But all these equidistributions have entropy zero.

### The Variational Principle: A Global Characterization of Equilibrium

Usually, a dynamical systems model of a “physical” system consists of a state space and a map (or a differential equation) describing the dynamics. An invariant measure for the system is rarely given a priori. Indeed, many (if not most) dynamical systems arising in this way have uncountably many ergodic invariant measures. This limits considerably the “practical value” of Birkhoff’s ergodic theorem (8) or the Shannon–McMillan–Breiman theorem (10): not only do the limits in these theorems depend on the invariant measure  $\nu$ , but also the sets of points for which the theorems guarantee almost everywhere convergence are practically disjoint for different  $\nu$  and  $\nu'$  in  $\mathcal{M}_\sigma(X)$ . Therefore, a choice of  $\nu$  has to be made which reflects the original modeling intentions. We will argue in this and the next sections that a variational principle with a judiciously chosen “observable” may be a useful guideline – generalizing the observations for finite systems collected in the corresponding section above. As announced earlier we restrict again to shift dynamical systems, because they are rather universal models for many other systems.

#### Equilibrium States

We define the *pressure* of an observable  $\phi \in C(X)$  as

$$P(\phi) := \sup\{h(\nu) + \langle \phi, \nu \rangle : \nu \in \mathcal{M}_\sigma(X)\}. \quad (12)$$

Since  $\mathcal{M}_\sigma(X)$  is compact and the functional  $\nu \mapsto h(\nu) + \langle \phi, \nu \rangle$  is upper semicontinuous, the supremum is attained – not necessarily at a unique measure as we will see

(which is remarkably different from what happens in finite systems). Each measure  $\nu$  for which the supremum is attained is called an *equilibrium state* for  $\phi$ . Here the word “state” is used synonymously with “distribution” or “measure” – a reflection of the fact that in “well-behaved cases”, as we will see in the next section, this measure is uniquely determined by the constraint(s) under which it maximizes entropy, and that means by the *macroscopic state* of the system. (In contrast, the word “state” was used in the above section on finite systems to designate microscopic states.)

As, for each  $\nu \in \mathcal{M}_\sigma(X)$ , the functional  $\phi \mapsto h(\nu) + \langle \phi, \nu \rangle$  is affine on  $C(X)$ , the pressure functional  $P: C(X) \rightarrow \mathbb{R}$ , which, by definition, is the pointwise supremum of these functionals, is convex. It is therefore instructive to fit equilibrium states into the abstract framework of convex analysis [32,38,45,68]. To this end recall the identities in (4) that identify, for finite systems, equilibrium states as gradients of the pressure function  $p: \mathbb{R}^{|A|} \rightarrow \mathbb{R}$  and guarantee that  $p$  is twice differentiable and strictly convex. In the present setting where  $P$  is defined on the Banach space  $C(X)$ , differentiability and strict convexity are no more guaranteed, but one can show:

#### Equilibrium states as (sub)-gradients

$\phi \in \mathcal{M}_\sigma(X)$  is an equilibrium state for  $\phi$  if and only if  $\mu$  is a subgradient (or tangent functional) for  $P$  at  $\phi$ , i. e., if  $P(\phi + \psi) - P(\phi) \geq \langle \psi, \mu \rangle$  for all  $\psi \in C(X)$ . In particular,  $\phi$  has a unique equilibrium state  $\mu$  if (13) and only if  $P$  is differentiable at  $\phi$  with gradient  $\mu$ , i. e., if  $\lim_{t \rightarrow 0} \frac{1}{t} (P(\phi + t\psi) - P(\phi)) = \langle \psi, \mu \rangle$  for all  $\psi \in C(X)$ .

Let us see how equilibrium states on  $X = A^\mathbb{N}$  can directly be obtained from the corresponding equilibrium distributions on finite sets  $A^n$  introduced in Subsect. “Systems on a Finite Lattice”. Define  $\phi^{(n)}: A^n \rightarrow \mathbb{R}$  by  $\phi^{(n)}(a_0 \dots a_{n-1}) := \phi(a_0 \dots a_{n-1} a_0 \dots a_{n-1} \dots)$ , denote by  $U_n$  the corresponding global observable on  $A^n$ , and let  $\mu_n$  be the equilibrium distribution on  $A^n$  that maximizes  $H(\mu) + \langle U_n, \mu \rangle$ . Then all weak limit points of the “approximative equilibrium distributions”  $\mu_n$  on  $A^n$  are equilibrium states on  $A^\mathbb{N}$ .

This can be seen as follows: Let the measure  $\mu$  on  $A^\mathbb{N}$  be any weak limit point of the  $\mu_n$ . Then, given  $\epsilon > 0$  there exists  $k \in \mathbb{N}$  such that

$$\begin{aligned} h(\mu) + \langle \phi, \mu \rangle &\geq \frac{1}{k} H_k(\mu) + \langle \phi, \mu \rangle - \epsilon \\ &\geq \frac{1}{k} H_k(\mu_n) + \langle \phi^{(n)}, \mu_n \rangle - 2\epsilon \end{aligned}$$

for arbitrarily large  $n$ , because  $\|\phi - \phi^{(n)}\|_\infty \rightarrow 0$  as  $n \rightarrow \infty$  by construction of the  $\phi^{(n)}$ . As the  $\mu_n$  are invariant under cyclic coordinate shifts (see Subsect. “[Systems on a Finite Lattice](#)”), it follows from the subadditivity of the entropy that

$$h(\mu) + \langle \phi, \mu \rangle \geq \frac{1}{n}(H_n(\mu_n) + \langle U_n, \mu_n \rangle) - 2\epsilon - \frac{k}{n} \log |A|.$$

Hence, for each  $v \in \mathcal{M}_\sigma(X)$ ,

$$h(\mu) + \langle \phi, \mu \rangle \geq \frac{1}{n}(H_n(v) + \langle U_n, v \rangle) - 2\epsilon - \frac{k}{n} \log |A| \rightarrow h(v) + \langle \phi, v \rangle - 2\epsilon$$

as  $n \rightarrow \infty$ , and we see that  $\mu$  is indeed an equilibrium state on  $A^\mathbb{N}$ .

### The Variational Principle

In Subsect. “[Equilibrium Distributions and the Gibbs Property](#)”, the pressure of a finite system was defined as a certain supremum and then identified as the logarithm of the normalizing constant for the Gibbsian representation of the corresponding equilibrium distribution. We are now going to approximate equilibrium states by suitable Gibbs distributions on finite subsets of  $X$ . As a by-product the pressure  $P(\phi)$  is characterized in terms of the logarithms of the normalizing constants of these approximating distributions. Let  $S_n\phi(\underline{x}) := \phi(\underline{x}) + \phi(\sigma\underline{x}) + \dots + \phi(\sigma^{n-1}\underline{x})$ . From each cylinder set  $[a_0 \dots a_{n-1}]$  we can pick a point  $\underline{z}$  such that  $S_n\phi(\underline{z})$  is the maximal value of  $S_n\phi$  on this set. We denote the collection of the  $|A|^n$  points we obtain in this way by  $E_n$ . Observe that  $E_n$  is not unambiguously defined, but any choice we make will do.

### Theorem (Variational Principle for the Pressure)

$$P(\phi) = \limsup_{n \rightarrow \infty} \frac{1}{n} P_n(\phi) \quad \text{where } P_n(\phi) := \log \sum_{\underline{z} \in E_n} e^{S_n\phi(\underline{z})}. \quad (14)$$

To prove the “ $\leq$ ” direction of this identity we just have to show that  $\frac{1}{n}H_n(v) + \langle \phi, v \rangle \leq \frac{1}{n}P_n(\phi)$  for each  $v \in \mathcal{M}_\sigma(X)$  or, after multiplying by  $n$ ,  $H_n(v) + \langle S_n\phi, v \rangle \leq$

$P_n(\phi)$ . But Jensen’s inequality implies:

$$\begin{aligned} H_n(v) + \langle S_n\phi, v \rangle &\leq \sum_{a_0, \dots, a_{n-1} \in A} v[a_0 \dots a_{n-1}] \\ &\quad \log \left( \frac{\sup \{e^{S_n\phi(\underline{x})} : \underline{x} \in [a_0 \dots a_{n-1}]\}}{v[a_0 \dots a_{n-1}]} \right) \\ &\leq \log \sum_{a_0, \dots, a_{n-1} \in A} \sup \{e^{S_n\phi(\underline{x})} : \underline{x} \in [a_0 \dots a_{n-1}]\} \\ &= \log \sum_{\underline{z} \in E_n} e^{S_n\phi(\underline{z})} = P_n(\phi). \end{aligned}$$

For the reverse inequality consider the discrete Gibbs distributions

$$\pi_n := \sum_{\underline{z} \in E_n} \delta_{\underline{z}} \exp(-P_n(\phi) + S_n\phi(\underline{z}))$$

on the finite sets  $E_n$ , where  $\delta_{\underline{z}}$  denotes the unit point mass in  $\underline{z}$ . One might be tempted to think that all weak limit points of the measures  $\pi_n$  are already equilibrium states. But this need not be the case because there is no good reason that these limits are shift invariant. Therefore, one forces invariance of the limits by passing to measures  $\mu_n$  defined by  $\langle f, \mu_n \rangle := \langle \frac{1}{n} \sum_{k=0}^{n-1} f \circ \sigma^k, \pi_n \rangle$ . Weak limits of these measures are obviously shift invariant, and a more involved estimate we do not present here shows that each such weak limit  $\mu$  satisfies  $h(\mu) + \langle \phi, \mu \rangle \geq P(\phi)$ .

We note that the same arguments work for any other sequence of sets  $E_n$  which contain exactly one point from each cylinder. So there are many ways to approximate equilibrium states, and if there are more than one equilibrium state, there is generally no guarantee that the limit is always the same.

### Nonuniqueness of Equilibrium States: An Example

Before we turn to sufficient conditions for the uniqueness of equilibrium states in the next section, we present one of the simplest nontrivial examples for nonuniqueness of equilibrium states. Motivated by the so-called Fisher–Felderhof droplet model of condensation in statistical mechanics [23,25], Hofbauer [31] studies an observable  $\phi$  on  $X = \{0, 1\}^\mathbb{N}$  defined as follows: Let  $(a_k)$  be a sequence of negative real numbers with  $\lim_{k \rightarrow \infty} a_k = 0$ . Set  $s_k := a_0 + \dots + a_k$ . For  $k \geq 1$  denote  $M_k := \{\underline{x} \in X : x_0 = \dots = x_{k-1} = 1, x_k = 0\}$  and  $M_0 := \{\underline{x} \in X : x_0 = 0\}$ , and define

$$\phi(\underline{x}) := a_k \quad \text{for } \underline{x} \in M_k \text{ and } \phi(11\dots) = 0.$$



Then  $\phi: X \rightarrow \mathbb{R}$  is continuous, so that there exists at least one equilibrium state for  $\phi$ . Hofbauer proves that there is more than one equilibrium state if and only if  $\sum_{k=0}^{\infty} e^{s_k} = 1$  and  $\sum_{k=0}^{\infty} (k+1)e^{s_k} < \infty$ . In that case  $P(\phi) = 0$ , so one of these equilibrium states is the unit mass  $\delta_{11\dots}$ , and we denote the other equilibrium state by  $\mu_1$ , so  $h(\mu_1) + \langle \phi, \mu_1 \rangle = 0$ . In view of (13) the pressure function is not differentiable at  $\phi$ .

What does the pressure function  $\beta \mapsto P(\beta\phi)$  look like? As  $h(\delta_{11\dots}) + \langle \beta\phi, \delta_{11\dots} \rangle = 0$  for all  $\beta$ ,  $P(\beta\phi) \geq 0$  for all  $\beta$ . Observe now that  $\phi(\underline{x}) \leq 0$  with equality only for  $\underline{x} = 11\dots$ . This implies that  $\langle \phi, \mu \rangle < 0$  for all  $\mu \in \mathcal{M}_\sigma(X)$  different from  $\delta_{11\dots}$ . From this we can conclude:

- $P(\beta\phi) \leq P(\phi) = 0$  for  $\beta > 1$ , so  $P(\beta\phi) = 0$  for  $\beta \geq 1$ .
- $P(\beta\phi) \geq h(\mu_1) + \langle \beta\phi, \mu_1 \rangle = h(\mu_1) + \langle \phi, \mu_1 \rangle - (1 - \beta)\langle \phi, \mu_1 \rangle = -(1 - \beta)\langle \phi, \mu_1 \rangle$ .

It follows that, at  $\beta = 1$ , the derivative from the right of  $P(\beta\phi)$  is zero, whereas the derivative from the left is at most  $-\langle \phi, \mu_1 \rangle < 0$ .

### More on Equilibrium States

In more general dynamical systems the entropy function is not necessarily upper semicontinuous and hence equilibrium states need not exist, i. e., the supremum in (12) need not be attained by any invariant measure. A well-known sufficient property that guarantees the upper semicontinuity of the entropy function is the *expansiveness* of the system, see, e. g., [53]: a continuous transformation  $T$  of a compact metric space is *positively expansive*, if there is a constant  $\gamma > 0$  such that for any two points  $x$  and  $y$  from the space there is some  $n \in \mathbb{N}$  such that  $T^n x$  and  $T^n y$  are at least a distance  $\gamma$  apart. If  $T$  is a homeomorphism one says it is *expansive*, if the same holds for some  $n \in \mathbb{Z}$ . The previous results carry over without changes (although at the expense of more complicated proofs) to general expansive systems. The variational principle (14) holds in the very general context where  $T$  is a continuous action of  $\mathbb{Z}_+^d$  on a compact Hausdorff space  $X$ . This was proved in [44] in a simple and elegant way. In the monograph [45] it is extended to amenable group actions.

### The Gibbs Property: A Local Characterization of Equilibrium

In this section we are going to see that, for a sufficiently regular potential  $\phi$  on a topologically mixing subshift of finite type, one has a unique equilibrium state which has the

“Gibbs property”. This property generalizes formula (5) that we derived for finite lattices. Subshifts of finite type are the symbolic models for Axiom A diffeomorphisms, as we shall see later on.

### Subshifts of Finite Type

We start by recalling what is a subshift of finite type and refer the reader to ► [Symbolic Dynamics](#) or [43] for more details. Given a “transition matrix”  $M = (M_{ab})_{a,b \in A}$  whose entries are 0’s or 1’s, one can define a subshift  $X_M$  as the set of all sequences  $\underline{x} \in A^{\mathbb{N}}$  such that  $M_{x_i x_{i+1}} = 1$  for all  $i \in \mathbb{N}$ . This is called a subshift of finite type or a topological Markov chain. We assume that there exists some integer  $p_0$  such that  $M^{p_0}$  has strictly positive entries for all  $p \geq p_0$ . This means that  $M$  is irreducible and aperiodic. This property is equivalent to the property that the subshift of finite type is topologically mixing. A general subshift of finite type admits a decomposition into a finite union of transitive sets, each of which being a union of cyclically permuted sets on which the appropriate iterate is topologically mixing. In other words, topologically mixing subshifts of finite type are the building blocks of subshifts of finite type.

### The Gibbs Property for a Class of Regular Potentials

The class of regular potentials we consider is that of “summable variations”. We denote by  $\text{var}_k(\phi)$  the modulus of continuity of  $\phi$  on cylinders of length  $k \geq 1$ , that is,

$$\text{var}_k(\phi) := \sup\{|\phi(\underline{x}) - \phi(\underline{y})| : \underline{x} \in [y_0 \dots y_{k-1}]\}.$$

If  $\text{var}_k(\phi) \rightarrow 0$  as  $k \rightarrow \infty$ , this means that  $\phi$  is (uniformly) continuous with respect to the distance (7). We impose the stronger condition

$$\sum_{k=1}^{\infty} \text{var}_k(\phi) < \infty. \quad (15)$$

We can now state the main result of this section.

**The Gibbs state of a summable potential** Let  $X_M$  be a topologically mixing subshift of finite type. Given a potential  $\phi: X_M \rightarrow \mathbb{R}$  satisfying the summability condition (15), there is a (probability) measure  $\mu_\phi$  supported on  $X_M$ , that we call a Gibbs state. It is the unique  $\sigma$ -invariant measure which satisfies the following property:

There exists a constant  $C > 0$  such that, for all  $\underline{x} \in X_M$  and for all  $n \geq 1$ ,

$$C^{-1} \leq \frac{\mu_\phi[x_0 \dots x_{n-1}]}{\exp(S_n \phi(\underline{x}) - nP(\phi))} \leq C. \quad (\text{“Gibbs property”}) \quad (16)$$

Moreover, the Gibbs state  $\mu_\phi$  is ergodic and is also the unique equilibrium state of  $\phi$ , i. e., the unique invariant measure for which the supremum in (12) is attained.

We now make several comments on this theorem.

- The Gibbs property (16) gives a *uniform control* of the measure of *all cylinders* in terms of their “energy”. This strengthens considerably the asymptotic equipartition property (11) that we recover if we restrict (16) to the set of  $\mu_\phi$  measure 1 where Birkhoff’s ergodic Theorem (8) applies, and use the identity  $\langle \phi, \mu_\phi \rangle - P(\phi) = -h(\mu_\phi)$ .
- Gibbs measures on topologically mixing subshifts of finite type are ergodic (and actually mixing in a strong sense) as can be inferred from Ruelle’s Perron–Frobenius Theorem see the next subsection.
- Suppose that there is another invariant measure  $\mu'$  satisfying (16), possibly with a constant  $C'$  different from  $C$ . It is easy to verify that  $\mu' = f\mu$  for some  $\mu$ -integrable function  $f$  by using (16) and the Radon–Nikodym Theorem. Shift invariance imposes that,  $\mu$ -a. e.,  $f = f \circ \sigma$ . Then the ergodicity of  $\mu$  implies that  $f$  is a constant  $\mu$ -a. e., thus  $\mu' = \mu$ ; see [6].
- One could define a Gibbs state by saying that it is an invariant measure  $\mu$  satisfying (16) for a given continuous potential  $\phi$ . If one does so, it is simple to verify that such a  $\mu$  must also be an equilibrium state. Indeed, using (16), one can deduce that  $\langle \phi, \mu \rangle + h(\mu) \geq P(\phi)$ . The converse need not be true in general, see Subsect. “[More on Hofbauer’s Example](#)” below. But the summability condition (15) is indeed sufficient for the coincidence of Gibbs and equilibrium states. A proof of this fact can be found in [58] or [38].

### Ruelle’s Perron–Frobenius Theorem

The powerful tool behind the theorem in the previous subsection is a far-reaching generalization of the classical Perron–Frobenius theorem for irreducible matrices. Instead of a matrix, one introduces the so-called transfer operator, also called the “Perron–Frobenius operator” or “Ruelle’s operator”, which acts on a suitable Banach space of observables. It is D. Ruelle [52] who first introduced this operator in the context of one-dimensional lattice gases

with exponentially decaying interactions. In our context, this corresponds to Hölder continuous potentials: these are potentials satisfying  $\text{var}_k(\phi) \leq c\theta^k$  for some  $c > 0$  and  $\theta \in (0, 1)$ . A proof of “Ruelle’s Perron–Frobenius Theorem” can be found in [4,6]. It was then extended to potentials with summable variations in [67]. We refer to the book of V. Baladi [1] for a comprehensive account on transfer operators in dynamical systems.

We content ourselves to define the transfer operator and state Ruelle’s Perron–Frobenius Theorem. Let  $\mathcal{L} : C(X_M) \rightarrow C(X_M)$  be defined by

$$\begin{aligned} (\mathcal{L}f)(\underline{x}) &:= \sum_{\underline{y} \in \sigma^{-1}\underline{x}} e^{\phi(\underline{y})} f(\underline{y}) \\ &= \sum_{a \in A: M(a, x_0)=1} e^{\phi(a\underline{x})} f(a\underline{x}). \end{aligned}$$

(Obviously,  $a\underline{x} := ax_0x_1\dots$ )

**Theorem (Ruelle’s Perron–Frobenius Theorem)** *Let  $X_M$  be a topologically mixing subshift of finite type. Let  $\phi$  satisfy condition (15). There exist a number  $\lambda > 0$ ,  $h \in C(X_M)$ , and  $\nu \in \mathcal{M}(X)$  such that  $h > 0$ ,  $\langle h, \nu \rangle = 1$ ,  $\mathcal{L}h = \lambda h$ ,  $\mathcal{L}^* \nu = \lambda \nu$ , where  $\mathcal{L}^*$  is the dual of  $\mathcal{L}$ . Moreover, for all  $f \in C(X_M)$ ,*

$$\|\lambda^{-n} \mathcal{L}^n f - \langle f, \nu \rangle \cdot h\|_\infty \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

By using this theorem, one can show that  $\mu_\phi := h\nu$  satisfies (16) and  $\lambda = e^{P(\phi)}$ .

Let us remark that for potentials which are such that  $\phi(\underline{x}) = \phi(x_0, x_1)$  (i. e., potentials constant on cylinders of length 2),  $\mathcal{L}$  can be identified with a  $|A| \times |A|$  matrix and the previous theorem boils down to the classical Perron–Frobenius theorem for irreducible aperiodic matrices [63]. The corresponding Gibbs states are nothing but Markov chains with state space  $A$  (Chapter 3 in [29]). We shall take another point of view below (Subsect. “[Markov Chains over Finite Alphabets](#)”).

### Relative Entropy

We now define the relative entropy of an invariant measure  $\nu \in \mathcal{M}_\sigma(X_M)$  given a Gibbs state  $\mu_\phi$  as follows. We first define

$$\begin{aligned} H_n(\nu|\mu_\phi) &:= \sum_{a_0, \dots, a_{n-1} \in A} \nu[a_0 \dots a_{n-1}] \log \frac{\nu[a_0 \dots a_{n-1}]}{\mu_\phi[a_0 \dots a_{n-1}]} \end{aligned} \quad (17)$$

with the convention  $0 \log(0/0) = 0$ . Now the relative entropy of  $\nu$  given  $\mu_\phi$  is defined as

$$h(\nu|\mu_\phi) := \limsup_{n \rightarrow \infty} \frac{1}{n} H_n(\nu|\mu_\phi).$$

(By applying Jensen's inequality, one verifies that  $h(\nu|\mu_\phi) \geq 0$ .) In fact the limit exists and can be computed quite easily using (16):

$$h(\nu|\mu_\phi) = P(\phi) - \langle \phi, \nu \rangle - h(\nu). \quad (18)$$

To prove this formula, we first make the following observation. It can be easily verified that the inequalities in (16) remain the same when  $S_n \phi$  is replaced by the “locally averaged” energy  $\tilde{\phi}_n := (v[x_0 \dots x_{n-1}])^{-1} \int_{[x_0 \dots x_{n-1}]} S_n \phi(y) d\nu(y)$  for any cylinder with  $v[x_0 \dots x_{n-1}] > 0$ . Cylinders with  $\nu$  measure zero does not contribute to the sum in (17).

We can now write that

$$\begin{aligned} & -\frac{1}{n} \log C \\ & \leq -\frac{1}{n} H_n(\nu|\mu_\phi) + \left( P(\phi) - \frac{1}{n} \langle S_n \phi, \nu \rangle - \frac{1}{n} H_n(\nu) \right) \\ & \leq \frac{1}{n} \log C. \end{aligned}$$

To finish we use that  $\langle S_n \phi, \nu \rangle = n \langle \phi, \nu \rangle$  (by the invariance of  $\nu$ ) and we apply (9) to obtain

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} H_n(\nu|\mu_\phi) &= P(\phi) - \langle \phi, \nu \rangle - \lim_{n \rightarrow \infty} \frac{1}{n} H_n(\nu) \\ &= P(\phi) - \langle \phi, \nu \rangle - h(\nu) \end{aligned}$$

which proves (18).

**The variational principle revisited** We can reformulate the variational principle in the case of a potential satisfying the summability condition (15):

$$h(\nu|\mu_\phi) = 0 \quad \text{if and only if } \nu = \mu_\phi, \quad (19)$$

i.e., given  $\mu_\phi$ , the relative entropy  $h(\cdot|\mu_\phi)$ , as a function on  $\mathcal{M}_\sigma(X_M)$ , attains its minimum only at  $\mu_\phi$ .

Indeed, by (18) we have  $h(\nu|\mu_\phi) = P(\phi) - \langle \phi, \nu \rangle - h(\nu)$ . We now use (12) and the fact that  $\mu_\phi$  is the unique equilibrium state of  $\phi$  to conclude.

### More Properties of Gibbs States

Gibbs states enjoy very good statistical properties. Let us mention only a few. They satisfy the “Bernoulli property”, a very strong qualitative mixing condition [4,6,67]. The sequence of random variables  $(f \circ \sigma^n)_n$  satisfies the central

limit theorem [15,16,49] and a large deviation principle if  $f$  is Hölder continuous [21,38,39,70]. Let us emphasize the central role played by relative entropy in large deviations. (The deep link between thermodynamics and large deviations is described in [42] in a much more general context.) Finally, the so-called “multifractal analysis” can be performed for Gibbs states, see, e.g., [48].

### Examples on Shift Spaces

#### Measure of Maximal Entropy and Periodic Points

If the observable  $\phi$  is constant zero, an equilibrium state simply maximizes the entropy. It is called *measure of maximal entropy*. The quantity  $P(0) = \sup\{h(\nu) : \nu \in \mathcal{M}_\sigma(X)\}$  is called the *topological entropy* of the subshift  $\sigma : X \rightarrow X$ . When  $X$  is a subshift of finite type  $X_M$  with irreducible and aperiodic transition matrix  $M$ , there is a unique measure of maximal entropy, see, e.g., [43]. As a Gibbs state it satisfies (16). By summing over all cylinders  $[x_0 \dots x_{n-1}]$  allowed by  $M$ , it is easy to see that the topological entropy  $P(0)$  is the asymptotic exponential growth rate of the number of sequences of length  $n$  that can occur as initial segments of points in  $X_M$ . This is obviously identical to the logarithm of the largest eigenvalue of the transition matrix  $M$ .

It is not difficult to verify that the total number of periodic sequences of period  $n$  equals the trace of the matrix  $M^n$ , i.e., we have the formula

$$\text{Card} \{ \underline{x} \in X_M : \sigma^n \underline{x} = \underline{x} \} = \text{tr}(M^n) = \sum_{i=1}^m \lambda_i^n,$$

where  $\lambda_1, \dots, \lambda_m$  are all the eigenvalues of  $M$ . Asymptotically, of course,  $\text{Card} \{ \underline{x} \in X_M : \sigma^n \underline{x} = \underline{x} \} = e^{nP(0)} + O(|\lambda'|^n)$ , where  $\lambda'$  is the second largest (in absolute value) eigenvalue of  $M$ .

The measure of maximal entropy, call it  $\mu_0$ , describes the distribution of periodic points in  $X_M$ : one can prove [3,37] that for any cylinder  $B \subset X_M$

$$\lim_{n \rightarrow \infty} \frac{\text{Card} \{ \underline{x} \in B : \sigma^n \underline{x} = \underline{x} \}}{\text{Card} \{ \underline{x} \in X_M : \sigma^n \underline{x} = \underline{x} \}} = \mu_0(B).$$

In other words, the finite atomic measures that assign equal weights  $1/\text{Card} \{ \underline{x} \in X_M : \sigma^n \underline{x} = \underline{x} \}$  to each periodic point in  $X_M$  with period  $n$  weakly converges to  $\mu_0$ , as  $n \rightarrow \infty$ . Each such measure has zero entropy while  $h(\mu_0) = P(0) > 0$ , so the entropy is not continuous on the space of invariant measures. It is, however, upper-semicontinuous (see Subsect. “Entropy as a Function of the Measure”).

In fact, it is possible to approximate any Gibbs state  $\mu_\phi$  on  $X_M$  in a similar way by finite atomic measures on periodic orbits, if one assigns weights properly (see, e.g., Theorem 20.3.7 in [37]).

### Markov Chains over Finite Alphabets

Let  $Q = (q_{a,b})_{a,b \in A}$  be an irreducible stochastic matrix over the finite alphabet  $A$ . It is well known (see, e.g., [63]) that there exists a unique probability vector  $\pi$  on  $A$  that defines a stationary Markov measure  $\nu_Q$  on  $X = A^{\mathbb{N}}$  by  $\nu_Q[a_0 \dots a_{n-1}] = \pi_{a_0} q_{a_0 a_1} \dots q_{a_{n-2} a_{n-1}}$ . We are going to identify  $\nu_Q$  as the *unique Gibbs distribution*  $\mu \in \mathcal{M}_\sigma(X)$  that *maximizes entropy under the constraints*  $\mu[ab] = \mu[a]q_{ab}$ , i.e.,  $\langle \phi^{ab}, \mu \rangle = 0$  ( $a, b \in A$ ), where  $\phi^{ab} := \mathbb{1}_{[ab]} - q_{ab} \mathbb{1}_{[a]}$ . Indeed, as  $\mu$  is a Gibbs measure, there are  $\beta_{ab} \in \mathbb{R}$  ( $a, b \in A$ ) and constants  $P \in \mathbb{R}$ ,  $C > 0$  such that

$$C^{-1} \leq \frac{\mu[x_0 \dots x_{n-1}]}{\exp(\sum_{a,b \in A} \beta_{ab} \phi_n^{ab}(\underline{x}) - nP)} \leq C \quad (20)$$

for all  $\underline{x} \in A^{\mathbb{N}}$  and all  $n \in \mathbb{N}$ . Let  $r_{ab} := \exp(\beta_{ab} - \sum_{b' \in A} \beta_{ab'} q_{ab'} - P)$ . Then the denominator in (20) equals  $r_{x_0 x_1} \dots r_{x_{n-2} x_{n-1}}$ , and it follows that  $\mu$  is equivalent to the stationary Markov measure defined by the (non-stochastic) matrix  $(r_{ab})_{a,b \in A}$ . As  $\mu$  is ergodic,  $\mu$  is this Markov measure, and as  $\mu$  satisfies the linear constraints  $\mu[ab] = \mu[a]q_{ab}$ , we conclude that  $\mu = \nu_Q$ .

### The Ising Chain

Here the task is to characterize all “spin chains” in  $\underline{x} \in \{-1, +1\}^{\mathbb{N}}$  (or, more commonly,  $\{-1, +1\}^{\mathbb{Z}}$ ) which are as random as possible with the constraint that two adjacent spins have a prescribed probability  $p \neq \frac{1}{2}$  to be identical. With  $\phi(\underline{x}) := x_0 x_1$  this is equivalent to requiring that  $\underline{x}$  is typical for a Gibbs distribution  $\mu_{\beta\phi}$  where  $\beta = \beta(p)$  is such that  $\langle \phi, \mu_{\beta\phi} \rangle = 2p - 1$ . It follows that there is a constant  $C > 0$  such that for each  $n \in \mathbb{N}$  and any two “spin patterns”  $\underline{a} = a_0 \dots a_{n-1}$  and  $\underline{b} = b_0 \dots b_{n-1}$

$$\left| \log \frac{\mu_{\beta\phi}[a_0 \dots a_{n-1}]}{\mu_{\beta\phi}[b_0 \dots b_{n-1}]} - \beta(N_{\underline{a}} - N_{\underline{b}}) \right| \leq C,$$

where  $N_{\underline{a}}$  and  $N_{\underline{b}}$  are the numbers of identical adjacent spins in  $\underline{a}$  and  $\underline{b}$ , respectively.

### More on Hofbauer’s Example

We come back to the example described in Subsect. “Nonuniqueness of Equilibrium States: An Example”. It is easy to verify that in that example  $\text{var}_{k+1}(\phi) = |a_k|$ .

For instance, if  $a_k = -1/(k+1)^2$  there is a unique Gibbs/equilibrium state. If  $a_k = -3 \log((k+1)/k)$  for  $k \geq 1$  and  $a_0 = -\log \sum_{j=1}^{\infty} j^{-3}$ , then from [31] we know that  $\phi$  admits more than one equilibrium state, one of them being  $\delta_{11\dots}$ , which cannot be a Gibbs state for any continuous  $\phi$ .

### Examples from Differentiable Dynamics

In this section we present a number of examples to which the general theory developed above does not apply directly but only after a transfer of the theory from a symbolic space to a manifold. We restrict to examples where the results can be transferred because those aspects of the smooth dynamics we focus on can be studied as well on a shift dynamical system that is obtained from the original one via symbolic coding. (We do not discuss the coding process itself which is sometimes far from trivial, but we focus on the application of the Gibbs and equilibrium theory.) There are alternative approaches where instead of the results the *concepts and (partly) the strategies of proofs* are transferred to the smooth dynamical systems. This has led both to an extension of the range of possible applications of the theory and to a number of refined results (because some special features of smooth systems necessarily get lost by transferring the analysis to a completely disconnected metric space).

In the following examples,  $T$  denotes a (possibly piecewise) differentiable map of a compact smooth manifold  $M$ . Points on the manifold are denoted by  $u$  and  $v$ . In all examples there is a Hölder continuous coding map  $\pi: X \rightarrow M$  from a subshift of finite type  $X$  onto the manifold which respects the dynamics, i.e.,  $T \circ \pi = \pi \circ \sigma$ . This *factor map*  $\pi$  is “nearly” invertible in the sense that the set of points in  $M$  with more than one preimage under  $\pi$  has measure zero for all  $T$ -invariant measures we are interested in. Hence such measures  $\tilde{\mu}$  on  $M$  correspond unambiguously to shift invariant measures  $\mu = \tilde{\mu} \circ \pi^{-1}$ . Similarly observables  $\tilde{\phi}$  on  $M$  and  $\phi = \tilde{\phi} \circ \pi$  on  $X$  are related.

### Uniformly Expanding Markov Maps of the Interval

A transformation  $T$  on  $M := [0, 1]$  is called a *Markov map*, if there are  $0 = u_0 < u_1 < \dots < u_N = 1$  such that each restriction  $T|_{(u_{i-1}, u_i)}$  is strictly monotone,  $C^{1+r}$  for some  $r > 0$ , and maps  $(u_{i-1}, u_i)$  onto a union of some of these  $N$  monotonicity intervals. It is called *uniformly expanding* if there is some  $k \in \mathbb{N}$  such that  $\lambda := \inf_x |(T^k)'(x)| > 1$ . It is not difficult to verify that the symbolic coding of such a system leads to a topological Markov chain over the alphabet  $A = \{1, \dots, N\}$ . To simplify the discussion we assume that the transition ma-

trix  $M$  of this topological Markov chain is irreducible and aperiodic.

Our goal is to find a  $T$ -invariant measure  $\tilde{\mu}$  represented by  $\mu \in \mathcal{M}_\sigma(X_M)$  which minimizes the relative entropy to Lebesgue measure on  $[0, 1]$

$$h(\tilde{\mu}|m) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{a_0, \dots, a_{n-1} \in \{1, \dots, N\}} \mu[a_0 \dots a_{n-1}] \log \frac{\mu[a_0 \dots a_{n-1}]}{v_n[a_0 \dots a_{n-1}]},$$

where  $v_n[a_0 \dots a_{n-1}] := |I_{a_0 \dots a_{n-1}}|$ . (Recall that, without insisting on invariance, this would just be the Lebesgue measure itself.) The existence of the limit will be justified below – observe that  $m$  is not a Gibbs state as  $v$  is in Eq. (17). The argument rests on the simple observation (implied by the uniform expansion and the piecewise Hölder-continuity of  $T'$ ) that  $T$  has *bounded distortion*, i. e., that there is a constant  $C > 0$  such that for all  $n \in \mathbb{N}$ ,  $a_0 \dots a_{n-1} \in \{1, \dots, N\}^n$  and  $u \in I_{a_0 \dots a_{n-1}}$  holds

$$\begin{aligned} C^{-1} &\leq |I_{a_0 \dots a_{n-1}}| \cdot |(T^n)'(u)| \leq C, \text{ or, equivalently,} \\ C^{-1} &\leq \frac{|I_{a_0 \dots a_{n-1}}|}{\exp(S_n \tilde{\phi}(u))} \leq C, \end{aligned} \quad (21)$$

where  $\tilde{\phi}(u) := -\log |T'(u)|$ . (Observe the similarity between this property and the Gibbs property (16).) Assuming bounded distortion we have at once

$$\begin{aligned} h(\tilde{\mu}|m) &= \lim_{n \rightarrow \infty} \frac{1}{n} \left( -H_n(\mu) - \sum_{k=0}^{n-1} \langle \phi \circ \sigma^k, \mu \rangle \right) \\ &= -h(\mu) - \langle \phi, \mu \rangle, \end{aligned}$$

and minimizing this relative entropy just amounts to maximizing  $h(\mu) + \langle \phi, \mu \rangle$  for  $\phi = -\log |T'| \circ \pi$ . As the results on Gibbs distributions from Sect. “[The Gibbs Property: A Local Characterization of Equilibrium](#)” apply, we conclude that

$$C^{-1} \leq \frac{\mu[a_0 \dots a_{n-1}]}{|I_{a_0 \dots a_{n-1}}|} \leq C$$

for some  $C > 0$ . So the unique  $T$ -invariant measure  $\tilde{\mu}$  that minimizes the relative entropy  $h(\tilde{\mu}|m)$  is equivalent to Lebesgue measure  $m$ . (The existence of an invariant probability measure equivalent to  $m$  is well known, also without invoking entropy theory. It is guaranteed by a “Folklore Theorem” [33].)

## Interval Maps with an Indifferent Fixed Point

The presence of just one point  $x \in [0, 1]$  such that  $T'(x) = 1$  dramatically changes the properties of the system. A canonical example is the map  $T_\alpha : x \mapsto x(1+2^\alpha x^\alpha)$  if  $x \in [0, 1/2[$  and  $x \mapsto 2x - 1$  if  $x \in [1/2, 1]$ . We have  $T'(0) = 1$ , i. e., 0 is an indifferent fixed point. For  $\alpha \in [0, 1[$  this map admits an absolutely continuous invariant probability measure  $d\mu(x) = h(x)dx$ , where  $h(x) \sim x^{-\alpha}$  when  $x \rightarrow 0$  [66]. In the physics literature, this type of map is known as the “Manneville–Pomeau” map. It was introduced as a model of transition from laminar to intermittent behavior [50]. In [28] the authors construct a piecewise affine version of this map to study the complexity of trajectories (in the sense of Subsect. “[A Short Digression on Complexity](#)”). This gives rise to a countable state Markov chain. In [69] the close connection to the Fisher–Felderhof model and Hofbauer’s example (see Subsect. “[Nonuniqueness of Equilibrium States: An Example](#)”) was realized. We refer to [61] for recent developments and a list of references.

## Axiom A Diffeomorphisms, Anosov Diffeomorphisms, Sinai–Ruelle–Bowen Measures

The first spectacular application of the theory of Gibbs measures to differentiable dynamical systems was Sinai’s approach to Anosov diffeomorphisms via Markov partitions [64] that allowed one to code the dynamics of these maps into a subshift of finite type and to study their invariant measures by methods from equilibrium statistical mechanics [65] that had been developed previously by Dobrushin, Lanford, and Ruelle [17,18,19,20,41]. Not much later this approach was extended by Bowen [2] to Smale’s Axiom A diffeomorphisms (and to Axiom A flows by Bowen and Ruelle [7]); see also [54]. The interested reader can consult, e. g., [71] for a survey, and either [6] or [15] for details.

Both types of diffeomorphisms act on a smooth compact Riemannian manifold  $M$  and are characterized by the existence of a compact  $T$ -invariant *hyperbolic set*  $\Lambda \subseteq M$ . Their basic properties are described in detail in the contribution ▶ [Ergodic Theory: Basic Examples and Constructions](#). Very briefly, the tangent bundle over  $\Lambda$  splits into two invariant subbundles – a stable one and an unstable one. Correspondingly, through each point of  $\Lambda$  there passes a local stable and a local unstable manifold which are both tangent to the respective subspaces of the local tangent space. The unstable derivative of  $T$ , i. e., the derivative  $DT$  restricted to the unstable subbundle, is uniformly expanding. Its Jacobian determinant, denoted by  $J^{(u)}$ , is Hölder continuous as a function on  $\Lambda$ . Hence the



observable  $\phi^{(u)} := -\log |J^{(u)}| \circ \pi$  is Hölder continuous, and the Gibbs and equilibrium theory apply (via the symbolic coding) to the diffeomorphism  $T$  (modulo possibly a decomposition of the hyperbolic set into irreducible and aperiodic components, called basic sets, that can be modeled by topologically mixing subshifts of finite type). The main results are:

**Characterization of attractors** The following assertions are equivalent for a basic set  $\Omega \subseteq \Lambda$ :

- (i)  $\Omega$  is an attractor, i.e., there are arbitrarily small neighborhoods  $U \subseteq M$  of  $\Omega$  such that  $TU \subset U$ .
- (ii) The union of all stable manifolds through points of  $\Omega$  is a subset of  $M$  with positive volume.
- (iii) The pressure  $P_{T|_{\Omega}}(\phi^{(u)}) = 0$ .

In this case the unique equilibrium and Gibbs state  $\mu^+$  of  $T|_{\Omega}$  is called the *Sinai–Ruelle–Bowen (SRB) measure* of  $T|_{\Omega}$ . It is uniquely characterized by the identity  $h_{T|_{\Omega}}(\mu^+) = -\langle \phi^{(u)}, \mu^+ \rangle$ . (For all other  $T$ -invariant measures on  $\Omega$  one has “ $<$ ” instead of “ $=$ ”.)

**Further properties of SRB measures** Suppose  $P_{T|_{\Omega}}(\phi^{(u)}) = 0$  and let  $\mu^+$  be the SRB measure.

- (a) For a set of points  $u \in M$  of positive volume we have:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k u) = \langle f, \mu^+ \rangle.$$

(Indeed, because of (ii) of the above characterization, this holds for almost all points of the union of the stable manifolds through points of  $\Omega$ .)

- (b) Conditioned on unstable manifolds,  $\mu^+$  is absolutely continuous to the volume measure on unstable manifolds.

In the special case of transitive Anosov diffeomorphisms, the whole manifold is a hyperbolic set and  $\Omega = M$ . Because of transitivity, property (ii) from the characterization of attractors is trivially satisfied, so there is always a unique SRB measure  $\mu^+$ . As  $T^{-1}$  is an Anosov diffeomorphism as well – only the roles of stable and unstable manifolds are interchanged –  $T^{-1}$  has a unique SRB measure  $\mu^-$  which is the unique equilibrium state of  $T^{-1}$  (and hence also of  $T$ ) for  $\phi^{(s)} := \log |J^{(s)}|$ . One can show:

**SRB measures for Anosov diffeomorphisms** The following assertions are equivalent:

- (i)  $\mu^+ = \mu^-$ .
- (ii)  $\mu^+$  or  $\mu^-$  is absolutely continuous w.r.t the volume measure on  $M$ .

- (iii) For each periodic point  $u = T^n u \in M$ ,  $|J(u)| = 1$ , where  $J$  denotes the determinant of  $DT$ .

We remark that, similarly to the case of Markov interval maps, the unstable Jacobian of  $T^n$  at  $u$  is asymptotically equivalent to the volume of the “ $n$ -cylinder” of the Markov partition around  $u$ . So the maximization of  $h(\mu) + \langle \phi^{(u)}, \mu \rangle$  by the SRB measure  $\mu^+$  can again be interpreted as the minimization of the relative entropy of invariant measures with respect to the normalized volume, and the fact that  $P(\phi^{(u)}) = 0$  in the Anosov (or more generally attractor) case means that  $\mu^+$  is as close to being absolutely continuous as it is possible for a singular measure. This is reflected by the above properties (a) and (b).

We emphasize the meaning of property (a) above: it tells us that the SRB measure  $\mu^+$  is the only *physically observable* measure. Indeed, in numerical experiments with physical models, one picks an initial point  $u \in M$  “at random” (i.e., with respect to the volume or Lebesgue measure) and follows its orbit  $T^k u$ ,  $k \geq 0$ .

### Bowen’s Formula for the Hausdorff Dimension of Conformal Repellers

Just as nearby orbits converge towards an attractor, they diverge away from a repeller. Conformal repellers form a nice class of systems which can be coded by a subshift of finite type. The construction of their Markov partitions is much simpler than that of Anosov diffeomorphisms, see, e.g., [72].

Let us recall the definition of a conformal repeller before giving a fundamental example. Given a holomorphic map  $T: V \rightarrow \mathbb{C}$  where  $V \subset \mathbb{C}$  is open and  $J$  a compact subset of  $\mathbb{C}$ , one says that  $(J, V, T)$  is a conformal repeller if

- (i) there exist  $C > 0$ ,  $\alpha > 1$  such that  $|(T^n)'(z)| \geq C\alpha^n$  for all  $z \in J$ ,  $n \geq 1$ ;
- (ii)  $J = \bigcap_{n \geq 1} T^{-n}(V)$ , and
- (iii) for any open set  $U$  such that  $U \cap J \neq \emptyset$ , there exists  $n$  such that  $T^n(U \cap J) \supset J$ .

From the definition it follows that  $T(J) = J$  and  $T^{-1}(J) = J$ .

A fundamental example is the map  $T: z \rightarrow z^2 + c$ ,  $c \in \mathbb{C}$  being a parameter. It can be shown that for  $|c| < \frac{1}{4}$  there exists a compact set  $J$ , called a (hyperbolic) *Julia set*, such that  $(J, \mathbb{C}, T)$  is a conformal repeller.

Conformal repellers  $J$  are in general fractal sets and one can measure their “degree of fractality” by means of their Hausdorff dimension,  $\dim_H(J)$ . Roughly speaking, one computes this dimension by covering the set  $J$  by balls

with radius less than or equal to  $\delta$ . If  $N_\delta(J)$  denotes the cardinality of the smallest such covering, then we expect that

$$N_\delta(J) \sim \delta^{-\dim_H(J)}, \quad \text{as } \delta \rightarrow 0.$$

We refer the reader to ► [Ergodic Theory: Fractal Geometry](#) or [22,46] for a rigorous definition (based on Carathéodory's construction) and for more information on fractal geometry.

Bowen's formula relates  $\dim_H(J)$  to the unique zero of the pressure function  $\beta \mapsto P(\beta\tilde{\phi})$  where  $\tilde{\phi} := -(\log |T'|)|_J$ . It is not difficult to see that indeed this map has a unique zero for some positive  $\beta$ .

By property (i),  $S_n\tilde{\phi} \leq \text{const} - n \log \alpha$ , which implies (by (13)) that  $\frac{d}{d\beta} P(\beta\tilde{\phi}) = \langle \tilde{\phi}, \mu_\beta \rangle \leq -\log \alpha < 0$ . As  $P(0)$  equals the topological entropy of  $J$ , i. e., the logarithm of the largest eigenvalue of the matrix  $M$  associated to the Markov partition,  $P(0)$  is strictly positive. Therefore, (recall that the pressure function is continuous) there exists a unique number  $\beta_0 > 0$  such that  $P(\beta_0\tilde{\phi}) = 0$ .

It turns out that this unique zero is precisely  $\dim_H(J)$ :

**Bowen's formula** The Hausdorff dimension of  $J$  is the unique solution of the equation  $P(\beta\tilde{\phi}) = 0$ ,  $\beta \in \mathbb{R}$ ; in particular

$$P(\dim_H(J)\tilde{\phi}) = 0.$$

This formula was proven in [55] for a general class of conformal repellers after the seminal paper [5]. The main tool is a distortion estimate very similar to (21). A simple exposition can be found in [72].

### Nonequilibrium Steady States and Entropy Production

SRB measures for Anosov diffeomorphisms and Axiom A attractors have been accepted recently as conceptual models for *nonequilibrium steady states* in nonequilibrium statistical mechanics. Let us point out that the word “equilibrium” is used in physics in a much more restricted sense than in ergodic theory. Only diffeomorphisms preserving the natural volume of the manifold (or a measure equivalent to the volume) would be considered as appropriate toy models of physical equilibrium situations. In the case of Anosov diffeomorphisms this is precisely the case if the “forward” and “backward” SRB measures  $\mu^+$  and  $\mu^-$  coincide. Otherwise, the diffeomorphism models a situation out of equilibrium, and the difference between  $\mu^+$  and  $\mu^-$  can be related to entropy production and irreversibility.

Gallavotti and Cohen [26,27] introduced SRB measures as idealized models of nonequilibrium steady states

around 1995. In order to have as firm a mathematical basis as possible they made the “*chaotic hypothesis*” that the systems they studied behave like transitive Anosov systems. Ruelle [56] extended their approach to more general (even nonuniformly) hyperbolic dynamics; see also his reviews [57,59] for more recent accounts discussing also a number of related problems; see by [51], too. The importance of the Gibbs property of SRB measures for the discussion of entropy production was also highlighted in [35], where it is shown that for transitive Anosov diffeomorphisms the relative entropy  $h(\mu^+|\mu^-)$  equals the average entropy production rate  $\langle \log |J|, \mu^+ \rangle$  of  $\mu^+$  where  $J$  denotes again the Jacobian determinant of the diffeomorphism. In particular, the entropy production rate is zero if, and only if,  $h(\mu^+|\mu^-) = 0$ , i. e., using coding and (19), if, and only if,  $\mu^+ = \mu^-$ . According to Subsect. “[Axiom A Diffeomorphisms, Anosov Diffeomorphisms, Sinai–Ruelle–Bowen Measures](#)”, this is also equivalent to  $\mu^+$  or  $\mu^-$  being absolutely continuous with respect to the volume measure.

### Some Ongoing Developments and Future Directions

As we saw, many dynamical systems with uniform hyperbolic structure (e. g., Anosov maps, axiom A diffeomorphisms) can be modeled by subshifts of finite type over a finite alphabet. We already mentioned in Subsect. “[Interval Maps with an Indifferent Fixed Point](#)” the typical example of a map of the interval with an indifferent fixed point, whose symbolic model is still a subshift of finite type, but with a countable alphabet. The thermodynamic formalism for such systems is by now well developed [24,30,60,61,62] and used, e. g., for multidimensional piecewise expanding maps [13]. An active line of research is related to systems admitting representations by symbolic models called “towers” constructed by using “inducing schemes”. The fundamental example is the class of one-dimensional unimodal maps satisfying the “Collet–Eckmann condition”. A first attempt to develop thermodynamic formalism for such systems was made in [10] where existence and uniqueness of equilibrium measures for the potential function  $\tilde{\phi}_\beta(u) = -\beta \log |T'(u)|$  with  $\beta$  close to 1 was established. Very recently, new developments in this direction appeared, see, e. g., [11,12,47].

A largely open field of research concerns a new branch of nonequilibrium statistical mechanics, the so-called “chaotic scattering theory”, namely the analysis of chaotic systems with various openings or holes in phase space, and the corresponding repellers on which interesting invariant measures exist. We refer the reader to [15] for a brief account and references to the physics litera-

ture. The existence of (generalized) steady states on repellers and the so-called “escape rate formula” have been observed numerically in a number of models. So far, little has been proven mathematically, except for Anosov diffeomorphisms with special holes [15] and for certain nonuniformly hyperbolic systems [9].

## Bibliography

- Baladi V (2000) Positive transfer operators and decay of correlations. *Advanced Series in Nonlinear Dynamics*, vol 16. World Scientific, Singapore
- Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. *Amer J Math* 92:725–747
- Bowen R (1974/1975) Some systems with unique equilibrium states. *Math Syst Theory* 8:193–202
- Bowen R (1974/75) Bernoulli equilibrium states for Axiom A diffeomorphisms. *Math Syst Theory* 8:289–294
- Bowen R (1979) Hausdorff dimension of quasicircles. *Inst Hautes Études Sci Publ Math* 50:11–25
- Bowen R (2008) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. *Lecture Notes in Mathematics*, vol 470, 2nd edn (1st edn 1975). Springer, Berlin
- Bowen R, Ruelle D (1975) The ergodic theory of Axiom A flows. *Invent Math* 29:181–202
- Brudno AA (1983) Entropy and the complexity of the trajectories of a dynamical system. *Trans Mosc Math Soc* 2:127–151
- Bruin H, Demers M, Melbourne I (2007) Existence and convergence properties of physical measures for certain dynamical systems with holes. Preprint
- Bruin H, Keller G (1998) Equilibrium states for  $S$ -unimodal maps. *Ergodic Theory Dynam Syst* 18(4):765–789
- Bruin H, Todd M (2007) Equilibrium states for potentials with  $\sup \varphi - \inf \varphi < h_{\text{top}}(f)$ . *Commun Math Phys* doi:10.1007/s00220-0-008-0596-0
- Bruin H, Todd M (2007) Equilibrium states for interval maps: the potential  $-t \log |Df|$ . Preprint
- Buzzi J, Sarig O (2003) Uniqueness of equilibrium measures for countable Markov shifts and multidimensional piecewise expanding maps. *Ergod Theory Dynam Syst* 23(5):1383–1400
- Chaitin GJ (1987) Information, randomness & incompleteness. *Papers on algorithmic information theory*. World Scientific Series in Computer Science, vol 8. World Scientific, Singapore
- Chernov N (2002) Invariant measures for hyperbolic dynamical systems. In: *Handbook of Dynamical Systems*, vol 1A. North-Holland, pp 321–407
- Coelho Z, Parry W (1990) Central limit asymptotics for shifts of finite type. *Isr J Math* 69(2):235–249
- Dobrushin RL (1968) The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probab Appl* 13:197–224
- Dobrushin RL (1968) Gibbsian random fields for lattice systems with pairwise interactions. *Funct Anal Appl* 2:292–301
- Dobrushin RL (1968) The problem of uniqueness of a Gibbsian random field and the problem of phase transitions. *Funct Anal Appl* 2:302–312
- Dobrushin RL (1969) Gibbsian random fields. The general case. *Funct Anal Appl* 3:22–28
- Eizenberg A, Kifer Y, Weiss B (1994) Large deviations for  $\mathbb{Z}^d$ -actions. *Comm Math Phys* 164(3):433–454
- Falconer K (2003) *Fractal geometry. Mathematical foundations and applications*, 2nd edn. Wiley, San Francisco
- Fisher ME, Felderhof BU (1970) Phase transition in one-dimensional cluster interaction fluids: IA. Thermodynamics, IB. Critical behavior. II. Simple logarithmic model. *Ann Phys* 58:177–280
- Fiebig D, Fiebig U-R, Yuri M (2002) Pressure and equilibrium states for countable state Markov shifts. *Isr J Math* 131:221–257
- Fisher ME (1967) The theory of condensation and the critical point. *Physics* 3:255–283
- Gallavotti G (1996) Chaotic hypothesis: Onsager reciprocity and fluctuation-dissipation theorem. *J Stat Phys* 84:899–925
- Gallavotti G, Cohen EGD (1995) Dynamical ensembles in stationary states. *J Stat Phys* 80:931–970
- Gaspard P, Wang X-J (1988) Sporadicity: Between periodic and chaotic dynamical behaviors. In: *Proceedings of the National Academy of Sciences USA*, vol 85, pp 4591–4595
- Georgii H-O (1988) Gibbs measures and phase transitions. In: *de Gruyter Studies in Mathematics*, 9. de Gruyter, Berlin
- Gurevich BM, Savchenko SV (1998) Thermodynamic formalism for symbolic Markov chains with a countable number of states. *Russ Math Surv* 53(2):245–344
- Hofbauer F (1977) Examples for the nonuniqueness of the equilibrium state. *Trans Amer Math Soc* 228:223–241
- Israel R (1979) Convexity in the theory of lattice gases. *Princeton Series in Physics*. Princeton University Press
- Jakobson M, Świątek (2002) One-dimensional maps. In: *Handbook of Dynamical Systems*, vol 1A. North-Holland, Amsterdam, pp 321–407
- Jaynes ET (1989) *Papers on probability, statistics and statistical physics*. Kluwer, Dordrecht
- Jiang D, Qian M, Qian M-P (2000) Entropy production and information gain in Axiom A systems. *Commun Math Phys* 214:389–409
- Katok A (2007) Fifty years of entropy in dynamics: 1958–2007. *J Mod Dyn* 1(4):545–596
- Katok A, Hasselblatt B (1995) *Introduction to the modern theory of dynamical systems*. *Encyclopaedia of Mathematics and its Applications*, vol 54. Cambridge University Press, Cambridge
- Keller G (1998) Equilibrium states in ergodic theory. In: *London Mathematical Society Student Texts*, vol 42. Cambridge University Press, Cambridge
- Kifer Y (1990) Large deviations in dynamical systems and stochastic processes. *Trans Amer Math Soc* 321:505–524
- Kolmogorov AN (1983) Combinatorial foundations of information theory and the calculus of probabilities. *Uspekhi Mat Nauk* 38:27–36
- Lanford OE, Ruelle D (1969) Observables at infinity and states with short range correlations in statistical mechanics. *Commun Math Phys* 13:194–215
- Lewis JT, Pfister C-E (1995) Thermodynamic probability theory: some aspects of large deviations. *Russ Math Surv* 50:279–317
- Lind D, Marcus B (1995) *An introduction to symbolic dynamics and coding*. Cambridge University Press, Cambridge
- Misiurewicz M (1976) A short proof of the variational principle for a  $\mathbb{Z}_+^N$  action on a compact space. In: *International Conference on Dynamical Systems in Mathematical Physics* (Rennes, 1975), *Astérisque*, No. 40, Soc Math France, pp 147–157
- Moulin-Ollagnier J (1985) Ergodic Theory and Statistical Mechanics. In: *Lecture Notes in Mathematics*, vol 1115. Springer, Berlin

46. Pesin Y (1997) Dimension theory in dynamical systems. Contemporary views and applications. University of Chicago Press, Chicago
47. Pesin Y, Senti S (2008) Equilibrium Measures for Maps with Inducing Schemes. *J Mod Dyn* 2(3):1–31
48. Pesin Y, Weiss (1997) The multifractal analysis of Gibbs measures: motivation, mathematical foundation, and examples. *Chaos* 7(1):89–106
49. Pollicott M (2000) Rates of mixing for potentials of summable variation. *Trans Amer Math Soc* 352(2):843–853
50. Pomeau Y, Manneville P (1980) Intermittent transition to turbulence in dissipative dynamical systems. *Comm Math Phys* 74(2):189–197
51. Rondoni L, Mejía-Monasterio C (2007) Fluctuations in nonequilibrium statistical mechanics: models, mathematical theory, physical mechanisms. *Nonlinearity* 20(10):R1–R37
52. Ruelle D (1968) Statistical mechanics of a one-dimensional lattice gas. *Commun Math Phys* 9:267–278
53. Ruelle D (1973) Statistical mechanics on a compact set with  $\mathbb{Z}^v$  action satisfying expansiveness and specification. *Trans Amer Math Soc* 185:237–251
54. Ruelle D (1976) A measure associated with Axiom A attractors. *Amer J Math* 98:619–654
55. Ruelle D (1982) Repellers for real analytic maps. *Ergod Theory Dyn Syst* 2(1):99–107
56. Ruelle D (1996) Positivity of entropy production in nonequilibrium statistical mechanics. *J Stat Phys* 85:1–23
57. Ruelle D (1998) Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. *J Stat Phys* 95:393–468
58. Ruelle D (2004) Thermodynamic formalism: The mathematical structures of equilibrium statistical mechanics. Second edition. Cambridge Mathematical Library. Cambridge University Press, Cambridge
59. Ruelle D (2003) Extending the definition of entropy to nonequilibrium steady states. *Proc Nat Acad Sci* 100(6):3054–3058
60. Sarig O (1999) Thermodynamic formalism for countable Markov shifts. *Ergod Theory Dyn Syst* 19(6):1565–1593
61. Sarig O (2001) Phase transitions for countable Markov shifts. *Comm Math Phys* 217(3):555–577
62. Sarig O (2003) Existence of Gibbs measures for countable Markov shifts. *Proc Amer Math Soc* 131(6):1751–1758
63. Seneta E (2006) Non-negative matrices and Markov chains. In: Springer Series in Statistics. Springer
64. Sinai Ja G (1968) Markov partitions and C-diffeomorphisms. *Funct Anal Appl* 2:61–82
65. Sinai Ja G (1972) Gibbs measures in ergodic theory. *Russ Math Surv* 27(4):21–69
66. Thaler M (1980) Estimates of the invariant densities of endomorphisms with indifferent fixed points. *Isr J Math* 37(4):303–314
67. Walters P (1975) Ruelle's operator theorem and  $g$ -measures. *Trans Amer Math Soc* 214:375–387
68. Walters P (1992) Differentiability properties of the pressure of a continuous transformation on a compact metric space. *J Lond Math Soc* (2) 46(3):471–481
69. Wang X-J (1989) Statistical physics of temporal intermittency. *Phys Rev A* 40(11):6647–6661
70. Young L-S (1990) Large deviations in dynamical systems. *Trans Amer Math Soc* 318:525–543
71. Young L-S (2002) What are SRB measures, and which dynamical systems have them? Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays. *J Statist Phys* 108(5–6):733–754
72. Zinsmeister M (2000) Thermodynamic formalism and holomorphic dynamical systems. *SMF/AMS Texts and Monographs*, 2. American Mathematical Society

## Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation

MIE ICHIHARA<sup>1</sup>, TAKESHI NISHIMURA<sup>2</sup>

<sup>1</sup> Earthquake Research Institute, University of Tokyo, Tokyo, Japan

<sup>2</sup> Tohoku University, Sendai, Japan

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Elementary Processes in a Single-Bubble Dynamics](#)

[Bubbly Magma in an Elastic Rock as a Pressure Source](#)

[Acoustic Bubbles in Hydrothermal Systems](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Impulse** The word *impulse* is used in many areas in different ways. In classical mechanics, the impulse is the integral of force with respect to time. It is also used to refer to a fast-acting force, which is often idealized by a step function or a delta function. In this text, it is used to represent any functional form of pressure increase, either static or transient, which can generate observable signals.

**Magma, melt, liquid** Magma is a general name for molten rock. It is fluid but contains solid and gas inclusions in liquid matrix. The matrix in magma is silicate melt (which is often called just melt), and that in a hydrothermal system is water.

**Volatile** Volatile is compound in silicate melt. The major component is H<sub>2</sub>O, of which concentration is 1–5 wt% depending mainly on pressure and composition of the melt. It exsolves from melt and forms gas bubbles at relatively low pressure (ca. 100 MPa corresponding to the litho-static pressure around several kilo-meters). The second major component is CO<sub>2</sub>. Although its



concentration is usually several ppm, some kinds of melts may dissolve 3–30 wt% of CO<sub>2</sub> at several GPa.

**Long period seismic events** Long-period (or very long-period) seismic events are dominant in the period from about 1 s to more than a few tens of seconds. These signals at volcanoes are considered to be generated by interaction or resonance between volcanic fluid and the surrounding medium.

**Ground deformation** Ground deformation is often observed at volcanoes when magma chambers inflate or deflate. Such ground deformation is detected by geodetic measurements such as GPS, tilt or strain meters, and the deformations often continue for a few tens of minutes to days or even months.

**Magma chamber** A magma chamber is a storage system of molten magma. It is generally hard to detect, but is probably located at from a shallow depth (ca. 1 km) to a few tens of km beneath the volcanoes. The shape and size have not been confirmed yet, but it is usually assumed to be rather round and hundreds to thousands of meters in scale. A magma storage system which has a horizontal extent is called a sill, and one which has a vertical extent is called a dike.

**Rectified diffusion and rectified heat transfer** Rectified diffusion is a mechanism which can push dissolved volatiles into bubbles in a sound field. Bubbles take in more volatiles during expansion than they discharge during contraction, mainly because of the following two non-linear effects. Firstly, during expansion the bubble radius becomes larger so that the bubble surface is also larger than the surface during contraction. Secondly, radial bubble expansion tangentially stretches the diffusion layer and sharpens the radial gradient of the volatile concentration in the diffusion layer, so that the volatile flux into the bubble. The mechanism also works to push heat into bubbles and enhances evaporation in a liquid-vapor system. The rectified diffusion and heat transfer have been known and studied in mechanical and chemical engineering.

**Bubble collapse** When the bubble is compressed, oscillates, or loses its mass by diffusion or phase change, it contracts to a very small size and sometimes disappear. Bubble collapse indicates the contraction of a bubble and does not necessarily indicate its disappearance.

tial in generating a variety of volcanic activities. In particular, the gas phase is mechanically distinct from the other phases and plays important roles in the dynamic phenomena of volcanoes. When we work on volcanic activities, we are almost certainly confronted with physics problems associated with bubbles.

The roles of bubbles in volcanic activities have been investigated mainly in three aspects. Firstly, the nucleation, growth, and expansion of bubbles is considered to be the main force that brings the magma to the surface [62,88]. Secondly, a single bubble, if it is sufficiently large, may generate seismic waves when it rapidly expands or accelerates in the volcanic conduit [11,35,81,95], and may generate acoustic waves in the air when it oscillates or bursts at the magma surface [36,82,95,96]. Thirdly, the existence of bubbles can significantly reduce the sound velocity [16,

#### Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Table 1

##### List of important variables and constants

Notation	Unit	Definition
$C_{eq}$	–	Equilibrium volatile concentration (weight fraction) in the liquid
$c_l$	$m s^{-1}$	Sound speed in the liquid
$c_{pg}$	$J kg^{-1} K^{-1}$	Heat capacity of the gas at constant pressure
$c_{pl}$	$J kg^{-1} K^{-1}$	Heat capacity of the liquid at constant pressure
$K_g$	Pa	Effective bulk modulus of the bubble
$K_l$	Pa	Bulk modulus of the liquid
$L$	$J kg^{-1}$	Latent heat
$p_g$	Pa	Pressure in the bubble
$p_g$	Pa	Pressure in the liquid far from the bubble
$R$	m	Bubble radius
$S$	m	Outer radius of the cellular bubble
$T$	K	Temperature
$U$	$m s^{-1}$	Translational velocity of the bubble
$\gamma$	–	Specific heat ratio
$\eta_l$	Pa s	Liquid viscosity
$\kappa_{gl}$	$m^2 s^{-1}$	Diffusivity of the volatile in the liquid
$\kappa_T$	$m^2 s^{-1}$	Thermal diffusivity in the bubble
$\kappa_{Tl}$	$m^2 s^{-1}$	Thermal diffusivity in the liquid
$\tilde{\mu}$	Pa	Effective stiffness of the magma chamber
$\mu_l$	Pa	Shear elasticity of the liquid
$\rho_l$	$kg m^{-3}$	Liquid density
$\Sigma$	$m s^{-1}$	Surface tension
$\sigma_\infty$	Pa	Ambient stress change given to the magma chamber
$\tau$	s	Maxwell relaxation time
$\omega$	$rad s^{-1}$	Angular frequency

#### Definition of the Subject

A volcano consists of solids, liquids, gases, and intermediate materials of any two of these phases. Mechanical and thermo-dynamical interactions of these phases are essen-



42] and increase the attenuation and dispersion of the waves [15,30,44]. This effect is considered to be relevant to many spectral features of seismic waves and air-waves associated with volcanic activities [3,10,22,45].

Studies on bubble dynamics relevant to the volcanology are spread over many research fields and cannot be covered by a single paper. Good review papers and textbooks have already been published on bubble phenomena in sound fields [61,70,73,74] and on the nucleation and growth of bubbles in magma [62]. In this paper, we discuss several bubble dynamics phenomena selected from a particular point of view that the bubbly fluid works as an impulse generator. Here the term impulse means a pressure increase, either static or transient, which can generate any observable signal (e. g. earthquakes, ground deformations, airwaves, and an eruption itself). Especially, we focus on the processes that the impulse is excited by non-linear coupling between the internal processes of a bubbly fluid and an external perturbation. The importance of these processes have recently become noticed as a possible triggering mechanism of eruptions, earthquakes, and inflation of a volcano [57,64]. Although it is generally considered that stress perturbation caused by preceding events is important, exact mechanisms to generate a pressure increase, which is required to trigger the subsequent events, are yet under discussion. In the first place, factors controlling single bubble dynamics are summarized as the elementary processes in the bubbly fluid. Then two distinct liquid-bubble systems are considered, both of which are included in a volcano. The one is a body of bubbly magma confined in an elastic chamber, where elasticity of the chamber, melt viscosity, and gas diffusion are important. The other is a hydrothermal system, where bubble oscillation, evaporation, and heat transfer are important.

## Introduction

### Elementary Processes in Single-Bubble Dynamics

Radial motion of a single bubble interacting with ambient pressure perturbation is the elementary process controlling behaviors of the liquid-bubble mixtures. Although it appears quite simple, it contains various mechanisms in plenty. The great variety of behaviors of a single bubble has attracted many scientists, among whom is Leonardo da Vinci [76]. Nowadays, knowledge of the single bubble dynamics is used and studies are continued in many academic and industrial areas such as mechanical engineering, chemical engineering, medical science, and earth science.

Factors which may control the radial motion of a bubble are the pressure difference inside and outside the bub-

### Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Table 2

#### List of characteristic times

Notation	Equation	Definition
$\tau_c$	(2)	Inertia-controlled bubble collapse
$\tau_g$	(25)	Mass diffusion in the liquid around the bubble
$\tau_T$	(24)	Thermal diffusion in the bubble
$\tau_v <$	(11)	Viscosity-controlled bubble expansion
$\omega_o$	(9)	Natural frequency of a bubble

ble, inertia and stress associated with the deformation of the surrounding liquid, propagation of pressure waves, heat and mass transport, phase transition at the bubble wall, chemical reactions, relative translational motion between the bubble and the liquid, and so on. Because including all these mechanisms at the same time in order to calculate the behavior of a bubble is unrealistic, we need to make adequate simplification and assumptions. Each mechanism has its own characteristic time scale in which the effect is dominant (Table 2), and its own effect on the bubble dynamics. Knowing the individual time scales and features is important when we want to understand and simulate a certain phenomenon correctly and efficiently. A brief review of some representative mechanisms with linearized analyses are presented in Sect. “[Elementary Processes in a Single-Bubble Dynamics](#)” for this purpose. Based on the results, geophysical phenomena and proposed models are discussed in the latter sections.

### Bubbly Magma in an Elastic Rock as a Pressure Source

We consider a body of bubbly magma confined in an elastic rock. Pressure perturbations to the system are caused by a change of tectonic stress due to local earthquakes, surface unloading by dome collapse, passing seismic waves from a near or distant source, or depressurization of the chamber by degassing or magma leakage. Dynamic response of the system may be relevant to subsequent activities of the volcano as follows.

Nishimura [64] investigated pressure re-equilibration between the bubbles, the melt, and the surrounding elastic medium. It is assumed that the pressure of the system is suddenly decreased. After re-equilibration, the original magma pressure is partially or completely recovered or even exceeded, depending on the size of the bubbles, stiffness of the elastic container, and the confining pressure. His model is used to explain rapid pressurization of a magma chamber triggered by the lava-dome collapse at Soufriere Hills Volcano [97], and pressure recovery in

magma filling the chamber after explosive degassing to continue activities at Popocatepetl Volcano [12]. Shimomura et al. [87] extended the formulation of [64] to calculate the time profile of the pressure recovery after sudden decompression. They showed that the time scale of the pressure recovery is strongly controlled by the system parameters, which include stiffness of the elastic container, bubble number density, diffusivity of the volatile in the melt, ambient pressure, and properties of the melt. Chouet et al. [13] also calculated the time profile assuming the system parameters for Popocatepetl Volcano, and compared the results with a particular source time function of a very-long-period seismic signal. Furthermore, in the same year, Lensky et al. [53] independently developed a mathematically equivalent model considering magma with CO<sub>2</sub> bubbles in mantle rock. They interpreted the results as a possible pressurization mechanism to initiate dikes in mantle which allow the fast transport of magma. There are in fact quite a few documented cases in which eruptions were triggered by local tectonic earthquakes (e.g. [47,68]) and wave propagation from a distant earthquake (e.g. [6,54]). Recently, Manga and Brodsky [57] have given a comprehensive review on the phenomena and possible mechanisms. Brodsky et al. [6] investigated the possibility that a strain wave from a distant earthquake can increase the pressure in bubbly magma by rectified diffusion, which is the mechanism by which volatiles are pumped into a bubble by cyclic expansion and contraction. However, it has turned out that the mechanism by itself can cause a negligibly small pressure increase [6,29]. Several other mechanisms for long-range triggering have been proposed, which include pressure increase from rising bubbles [55], sub-critical crack growth [7], and fracture unclugging [8].

### Acoustic Bubbles in Hydrothermal Systems

A hydrothermal system is another major source of pressure increase, long-period volcano seismic events [46], and triggered seismicity [57,89]. Behaviors of a single bubble and liquid-bubble mixtures in a hydrothermal system are quite different from those in a magmatic system, mainly because of the water viscosity which is less than that of magma by several orders of magnitude. We introduce several phenomena which are particular to the hydrothermal systems in Sect. “Acoustic Bubbles in Hydrothermal Systems”.

Geyser activity is well known for intermittent activity of hot-water effusion. The effusion process looks quite similar to volcanic eruptions, and some geysers are characterized by regular intervals of time and duration, which are

also recognized in particular types of eruptions and seismic activities. Consequently, the geysers have been widely studied using seismological and geophysical techniques, as well as field observations, not only for clarifying the mechanism of the geysers but also for understanding the volcanic activities (e.g., [39,40,43,65]). Kedar et al. [39,40] conducted a unique experiment at Old Faithful Geyser, Yellowstone. They measured pressure within the geyser's water column simultaneously with seismic measurements on the surface. The data demonstrated that the tremor observed at Old Faithful results from impulsive events in the geyser. The impulsive events were modeled by a collapse of a spherical bubble by cooling that occurred when the water column reached a critical temperature. Their data are reviewed in Sect. “Acoustic Bubbles in Hydrothermal Systems” in relation to other studies on the dynamics of gas and vapor bubbles.

### Elementary Processes in a Single-Bubble Dynamics

#### Equation of Motion for the Bubble Radius

Motion of a bubble is in fact a fluid dynamical problem for the liquid surrounding the bubble. The simplest model describing the behavior of a bubble is based on three assumptions:

- (1) The bubble is spherical,
- (2) The liquid is incompressible, and
- (3) The motion is radial.

Using the basic equations of fluid mechanics, which are the continuity equation and the momentum equation, and the force balance at the bubble surface, the first equation of motion for the bubble radius was obtained by Rayleigh [80]:

$$\rho_l \left( R \ddot{R} + \frac{3}{2} \dot{R}^2 \right) = p(R) - p_1, \quad (1)$$

where  $R$  is the bubble radius,  $\rho_l$  is the liquid density,  $p(R)$  is the pressure in the liquid at the bubble surface, and  $p_1$  is the pressure in the liquid at a large distance from the bubble. Using Eq. (1), Rayleigh [80] solved the problem of the collapse of an empty cavity in a large body of liquid at a constant  $p_1$  and showed the characteristic collapse time is

$$\tau_c = R_0 \sqrt{\rho_l / p_1}. \quad (2)$$

The time  $\tau_c$  is called the Rayleigh collapse time and is one of the most important time scales in the bubble dynamics.

Plesset [69] extended Eq. (1) including the effects of surface tension and time-dependent pressure field, and

Proitsky [71] included the effect of viscosity. The generalized Rayleigh equation for bubble dynamics is known as the Rayleigh–Plesset equation. liquid viscosity and surface tension. The generalized Rayleigh equation for bubble dynamics is called the Rayleigh–Plesset equation [70]:

$$\rho_l \left( R\ddot{R} + \frac{3}{2}\dot{R}^2 \right) = p_g - p_l - 4\eta_l \frac{\dot{R}}{R} - \frac{2\Sigma}{R}, \quad (3)$$

where  $\eta_l$  is the liquid viscosity, and  $\Sigma$  is the surface tension. Equation (3) is valid for a Newtonian liquid under conditions of negligible mass exchange at the bubble surface. A further generalized equation to which these two restrictions do not apply is [72]:

$$\begin{aligned} \rho_l \left( R\dot{u}_l + \frac{3}{2}u_l^2 \right) - J \left[ 2u_l + J \left( \frac{1}{\rho_g} - \frac{1}{\rho_l} \right) \right] \\ = p_g - p_l + \int_R^\infty \frac{3\tau_{rr}}{r} dr - \frac{2\Sigma}{R}, \end{aligned} \quad (4)$$

where  $u_l$  is the radial liquid velocity at the bubble surface,  $J = \rho_l(u_l - \dot{R})$  is the outgoing mass flux through the bubble wall,  $\rho_g$  is the density of the gas in the bubble, and  $\tau_{rr}$  is the normal radial stress. When the interfacial mass flux  $J$  vanishes,  $u_l = \dot{R}$  as in the left-hand side of the original Eq. (3).

While the above equations consider a single bubble in an infinite melt, magmatic systems often contain bubbles with some finite spacing. Cellular models of packing which include a finite volume of melt in interaction with each bubble have been employed for closely spaced bubbles [79]. When the elementary cell is represented by a sphere with an outer radius of  $S$ , the equation corresponding to (3) is [76]:

$$\begin{aligned} \rho_l \left[ R\ddot{R} \left( 1 - \frac{R}{S} \right) + \frac{3}{2}\dot{R}^2 \left( 1 - \frac{4R}{3S} + \frac{1}{3}\frac{R^4}{S^4} \right) \right] \\ = p_g - p_l - 4\eta_l \frac{\dot{R}}{R} \left( 1 - \frac{R^3}{S^3} \right) - \frac{2\Sigma}{R}. \end{aligned} \quad (5)$$

Equation (5) agrees with Eq. (3) for  $S \rightarrow \infty$ .

When there is no transport of heat or mass between the liquid and the bubble, the pressure in the bubble is determined by the instantaneous bubble radius alone. Using the ideal gas approximation, we have

$$p_g R^{3\gamma} = p_{go} R_o^{3\gamma}, \quad (6)$$

where  $\gamma$  is the specific heat ratio, and the subscript  $o$  indicates the equilibrium value of the variable. Substituting Eq. (6) into Eq. (3) for  $p_g$  and linearizing the equation, we obtain a damped oscillator equation:

$$\ddot{X} + 2b_v \dot{X} + \omega_o^2 X = -\frac{p'_l}{\rho_l R_o^2}, \quad (7)$$

$$b_v = \frac{2\eta_l}{\rho_l R_o^2}, \quad (8)$$

$$\omega_o = \frac{1}{R_o} \sqrt{\frac{3\gamma p_{go} - 2\Sigma/R_o}{\rho_l}}, \quad (9)$$

where  $X$  and  $p'_l$  are defined by  $R = R_o(1 + X)$  and  $p_l = p_{go} - 2\Sigma/R_o + p'_l$ , respectively. Equation (7) is useful to see the characteristic behaviors of a bubble and their time scales. The resonant frequency of the bubble is  $\omega_o$  (rad s<sup>-1</sup>). When the second term in the left-hand side of Eq. (7) dominates the first one in the time scale of the resonant oscillation, namely when  $\omega_o < b_v$ , the resonant oscillation is damped. In the case of a gas bubble with a radius of 10<sup>-3</sup> m in magma ( $\rho_l = 2500$  kg m<sup>-3</sup>) at 10 MPa ( $p_{go} - 2\Sigma/R_o = 10^7$  Pa), the frequency ( $\omega_o/(2\pi)$ ) is about 20 kHz. The oscillation is damped when the viscosity is larger than 160 Pa s. This viscosity is relatively small for magma. According to these estimations, we see that the free oscillation of a bubble in magma is possible in the limited cases that the viscosity is small and the bubble is large. We also see that the bubble oscillation is easily excited in water which has a viscosity about 10<sup>-3</sup> Pa s.

### Liquid Rheology

The shear rheology of the liquid surrounding the bubble is one of the controlling factors for the bubble dynamics. According to experimental results, magma has viscoelastic nature, which is the most simply represented by a linear Maxwell model [99]. Then the normal radial stress  $\tau_{rr}$  in Eq. (4) is related to the corresponding strain rate  $\dot{\epsilon}_{rr}$  by

$$\tau_{rr} = \mu_l \int_0^t \exp\left(-\frac{t-t'}{\tau}\right) \dot{\epsilon}_{rr} dt', \quad (10)$$

where  $\mu_l$  is the shear elasticity and  $\tau$  is the relaxation time. In the limit of  $t \ll \tau$ , the Maxwell relation (10) is reduced to a linear elastic stress-strain relation as  $\tau_{rr} = \mu_l \epsilon_{rr}$ . While in the limit of  $t \gg \tau$ , it is reduced to a Newtonian viscous relation, that is a linear stress-strain rate relation as  $\tau_{rr} = \mu_l \tau \dot{\epsilon}_{rr}$ , where  $\mu_l \tau$  corresponds to the Newtonian viscosity  $\eta_l$ .

Fogler and Goddard [21] first used the viscoelastic relation (10) in the generalized Rayleigh–Plesset Eq. (4) without mass flux, and demonstrated that the influence of the viscoelastic effects on the radial motion of a bubble is characterized by a dimensionless parameter called the Deborah number  $De = \tau/\tau_c$ , which compares the relaxation time  $\tau$  and Rayleigh collapse time  $\tau_c$  defined in Eq. (2): the influence is large when  $\tau \gg \tau_c$ . Extending the formulation by Fogler and Goddard [21] to a cellular bubble, Ichihara [28] investigated its characteristic behaviors

in magmatic conditions. It is shown that the elastic oscillation of a bubble, which occurs in the case of  $\tau \gg \tau_c$ , is in a frequency of order of MHz and with very small displacement of the bubble wall because of the large shear modulus of the magma.

Change of the bubble radius in magma is mainly controlled by the viscosity except in magma with very small viscosity in which the bubble oscillation is possible. Therefore, in most of the studies for bubble growth in magma, effects of liquid inertia and viscoelasticity are not considered, and Eq. (3) or (5) for a Newtonian fluid is used, neglecting the left-hand side terms representing the inertia [2,62,79,88]. Barclay et al. [2] analytically solved the problem of the viscosity-controlled bubble expansion for instantaneous decompression, and showed the characteristic expansion time is

$$\tau_v = \frac{4\eta_l}{3p_l}, \quad (11)$$

where  $p_l$  is the pressure in the liquid. The time  $\tau_v$  is one of the most important time scales of the bubble dynamics in magma [2,30], while the Rayleigh collapse time  $\tau_c$ , which is controlled by the inertia, is important in low-viscosity fluids including hydrothermal systems.

Definition of the viscous expansion time corresponding to Eq. (11) is different depending on which problems and literature are being referenced. The time scale of the entire expansion of a bubble for instantaneous decompression is represented by Eq. (11) using the reduced pressure for  $p_l$  [2]. Volumetric oscillation of a bubble in an acoustic field is prevented by the viscous resistance if the period is shorter than  $\tau_v$ , in this case with the initial static pressure for  $p_l$  [30]. When the bubble expansion is driven by a constant gas pressure, which occurs at the initial stage of diffusion-drive gas expansion when the gas is efficiently supplied from the liquid, the bubble grows approximately as  $R \sim R_o \exp[t\Delta p/(4\eta_l)]$  [62,93], where  $\Delta p$  is the pressure difference. In this case,  $\tau_v = 4\eta_l/\Delta p$ . The last case is discussed again later in the section of mass transport.

### Liquid Compressibility

The effect of liquid compressibility on radial motion of a bubble was first considered in connection with underwater explosions [14,41]. Liquid compressibility allows energy transport as a pressure wave so that it causes radiation damping. Noting that the effect is considerable in the case of a violent oscillation or collapse of a bubble, several mathematical approaches were proposed to include the effect in the equation of bubble radius. According to mathematical and numerical studies by Prosperetti and

Lezzi [78], which compared the proposed equations, the following Keller's equation [41] is widely accepted as the most adequate form.

$$\begin{aligned} \rho_l \left[ \left( 1 - \frac{\dot{R}}{c_l} \right) R \ddot{R} + \frac{3}{2} \left( 1 - \frac{\dot{R}}{3c_l} \right) \dot{R}^2 \right] \\ = \left( 1 + \frac{\dot{R}}{c_l} + \frac{R}{c_l} \frac{d}{dt} \right) \left( p_g - p_l - 4\eta_l \frac{\dot{R}}{R} - \frac{2\Sigma}{R} \right), \end{aligned} \quad (12)$$

where  $c_l$  is the sound speed in the liquid. Although Prosperetti and Lezzi [78] further proposed to use the liquid enthalpy at the bubble wall instead of the pressure for the best accuracy, Eq. (12) is generally used in the literature.

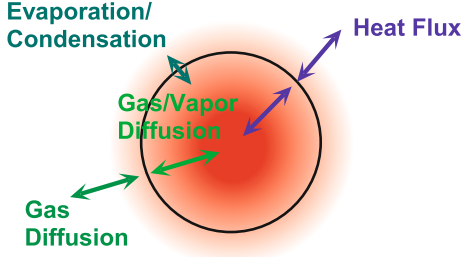
Comparing Eqs. (12) and (3), we can see that the correction terms due to the liquid compressibility have the order of  $\dot{R}/c_l$ . It means that the correction is considerable only when the bubble wall velocity becomes as large as the sound speed of the liquid. By applying  $[1 + (\dot{R}/c_l) + (R/c_l)d/dt]^{-1}$  to both sides of Eq. (12) and linearizing the equation, Prosperetti [75] derived the acoustic damping coefficient, which corresponds to  $b_v$  in Eq. (8) for the viscous damping, as  $b_{ac}$ :

$$b_{ac} = \frac{\omega^2 R_o}{2c_l}, \quad (13)$$

where  $\omega$  is the angular frequency of the oscillation. The acoustic damping is more significant when the bubble is larger and the oscillation frequency is higher.

Effects of liquid compressibility on bubble dynamics in a highly viscous liquid are not understood comprehensively. Derivation of Eq. (12) and related studies were performed thinking of liquids with ordinary viscosities like water. Therefore, the Reynolds number  $Re = \rho_l R_o c_l / \eta_l$  was presupposed to be large. On the other hand, the viscosity of magma can be large enough to make  $Re$  very small. In this case, the same mathematical approximation is not necessarily applicable. Yamada et al. [100] pointed out this problem and solved the equations including the viscous force associated with the volumetric strain rate. Although the equation of motion for the bubble radius appears not to be affected by the compressibility when the system is initially hydrostatic, the velocity field around the bubble is different from the incompressible solution, even if the wall velocity is much smaller than the acoustic velocity.

There is argument whether the equation of radial motion of a bubble surrounded by a finite volume of liquid needs correction terms for the compressibility. However, it seems to be negligible in magma, which is evaluated as follows [28]. In the case that the bubble is surrounded by an



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 1**

The internal heat and mass transport processes which control pressure change and energy loss associated with the bubble oscillation

elastic shell, contribution of the compressibility to the bubble expansion is  $\delta_c R = R_o(p_g - p_l)(1 - R_o^3/S_o^3)^{-1}(3K_l)^{-1}$ , where  $K_l$  is the bulk modulus of the liquid, which is the reciprocal of the compressibility [48]. We can evaluate  $\delta_c R/R_o < 10^{-3}$ , because  $K_l \sim 10^{10} - 10^{11}$  Pa for magma [99], the realistic pressure difference,  $p_g - p_l$  is not much larger than  $10^7$  Pa, and the volume fraction of the bubbles,  $R_o^3/S_o^3$ , is reasonably assumed to be smaller than the close-packing limit ( $\sim 0.74$ ). The change of  $p_g$  due to  $\delta_c R$  is  $\delta_c p_g/p_{go} \sim -3\delta_c R/R_o$ , which is also in the same order. When the shell deforms viscously, displacement due to non-volumetric deformation grows, while that from the volumetric deformation remains in the same order.

### Heat and Mass Transport

Each equation of motion for the bubble radius includes  $p_g$ , which is the pressure of the gas in the bubble, as we see in Eqs. (3), (5), and (12). Equation (6) is available to calculate  $p_g$  only for an adiabatic process. In fact, when a bubble expands, the pressure and temperature in the bubble decreases. Then heat and volatile components flow into the bubble from the surrounding liquid. The opposite occurs when a bubble shrinks. The internal processes which control  $p_g$  are schematically shown in Fig. 1. These transport effects are essential in most of actual systems including magmatic and hydrothermal systems.

Growth of a bubble by the mass diffusion in an over-saturated liquid is one of the fundamental problems. Based on purely dimensional considerations, an approximate growth law is given by

$$R\dot{R} = \frac{\kappa_{gl}\rho_l(C_o - C_{eq})}{\rho_g}, \quad (14)$$

where  $\kappa_{gl}$  is the diffusivity of the volatile in the liquid,  $C_o$  is the dissolved volatile concentration at a large distance

from the bubble, and  $C_{eq}$  is the equilibrium concentration at the given pressure [62,70]. From Eq. (14) we find that, asymptotically,  $R \sim \sqrt{2\kappa_{gl}\rho_l(C_o - C_{eq})t/\rho_g}$ . This expression is not valid at  $t \rightarrow 0$  making  $\dot{R} \rightarrow \infty$ . In the initial stage, the diffusion is very efficient and the bubble growth is controlled by viscous resistance [62,93]. In this stage,  $R \sim R_o \exp[t\Delta p/(4\eta_l)]$  as discussed in the section of liquid rheology. The approximate time of the transition from the viscosity-controlled exponential solution to the diffusion-controlled square-root solution is found by Navon and Lyakhovsky [62] to be

$$\tau_{vd} \sim [-15 - 10 \log(Pe)]\eta_l/\Delta p, \quad (15)$$

where  $Pe = \Delta p R_o^2 \eta_l^{-1} \kappa_{gl}^{-1}$  is the Pecret number that compares the time scales of viscous expansion and diffusion. It is noted that Eq. (15) is validated for  $Pe < 10^{-2}$ , that is for relatively large viscosity and small bubble radius [62]. If we consider  $\Delta p \sim 10^6$  Pa and  $\kappa_{gl} \sim 10^{-11} \text{ m}^2 \text{ s}^{-1}$ , this condition is satisfied when  $R_o^2 \eta_l^{-1} < 10^{-19}$ , that is  $\eta_l > 10^7$  Pa s when  $R_o \sim 10^{-6}$  m, and  $\eta_l > 10^9$  Pa s when  $R_o \sim 10^{-5}$  m. Then the corresponding times are  $\tau_{vd} > 50$  s and  $\tau_{vd} > 5000$  s, respectively. Lensky et al. [51] have suggested that the change of the characteristic behavior of the bubble expansion over the time scale  $\tau_{vd}$  generates a non-linear response of the liquid-bubble mixture to the pressure perturbation, which may cause amplification of a pressure wave. Coupling of effects of viscosity and diffusion on the bubble expansion also occurs through the material properties, because magma viscosity and water diffusivity are strongly influenced by the amount of dissolved water, which is the major volatile component in magma [4,50].

Matsumoto and Takemura [59] numerically solved a complete set of equations for the radial dynamics of a bubble including the conservation equation for mass, momentum, and energy in the bubble, heat and mass diffusion in the liquid, and heat and mass exchange between the gas and the liquid by diffusion and evaporation/condensation. Except in extremely rapid phenomena as the cases they treated, approximation of a spatially uniform pressure in the bubble is adequate [75]. With this approximation, the computational load is considerably reduced [9,38,63].

It is necessary to consider non-uniform temperature distribution and compositions, in order to quantify the amounts of energy exchange between the bubble and liquid and energy loss associated with the non-equilibrium process. Time scales required to recover uniform temperature and composition in the bubble are controlled by diffusion processes and are much longer than that to attain uniform pressure, which is controlled by the pressure



wave propagation in the bubble. Assuming representative values of the thermal diffusivity,  $\kappa_T \sim 10^{-5} \text{ (m}^2 \text{ s}^{-1}\text{)}$ , and the inter-diffusivity of the components in the gas phase,  $\kappa_{gi} \sim 10^{-7} \text{ (m}^2 \text{ s}^{-1}\text{)}$  [38], development of thermal and material diffusion layers all over the bubble with radius of  $10^{-3} \text{ m}$  takes  $\sim 0.1 \text{ s}$  and  $\sim 10 \text{ s}$ , respectively. The time range of  $0.1\text{--}10 \text{ s}$  is exactly what studies on seismoacoustic phenomena in volcanology are mainly concerned with. It takes an even longer amount of time to recover uniform concentration of volatile components in the liquid around the bubble. Therefore approximation of uniform temperature and compositions are not always adequate. Again we introduce results from the linearized theory for a periodic acoustic field. The bulk modulus of a bubble,  $K_g$ , is defined as:

$$K_g = -\frac{R}{3} \frac{\partial p_g}{\partial R}, \quad (16)$$

which is generally a complex number. The elasticity and the energy loss associated with volumetric change of a bubble are represented by the real and imaginary parts of  $K_g$ , respectively. Then the damping factor and the resonant frequency for the bubble oscillation, which correspond to Eqs. (8) and (9), respectively, are [75]:

$$b_t = \frac{3\text{Im}(K_g)}{2\rho_l \omega R_o^2} \quad (17)$$

$$\omega_o = \frac{1}{R_o} \sqrt{\frac{3\text{Re}(K_g) - 2\Sigma/R_o}{\rho_l}}. \quad (18)$$

In the case of an adiabatic process for an ideal gas, where Eq. (6) holds,  $K_g = \gamma p_{go}$  and Eq. (18) agree with Eq. (9). While in the case of an isothermal process,  $K_g = p_{go}$ . In these two extreme conditions,  $\text{Im}(K_g) = 0$  and there is no thermal damping.

In order to include the effect of non-uniform temperature distribution in the bubble, we have to solve the energy equation with the continuity of temperature at the bubble surface. Assuming that the pressure in the bubble is uniform, the temperature at the bubble wall is constant, which is supported by the large heat capacity of the liquid compared with that of the gas, and the pressure perturbation is periodic ( $\propto e^{i\omega t}$ ), the effective bulk modulus,  $K_g$ , is represented as [75]:

$$\frac{p_{go}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma} i\chi \left[ \sqrt{\frac{i}{\chi}} \coth \left( \sqrt{\frac{i}{\chi}} \right) - 1 \right], \quad (19)$$

$$\chi = \frac{\kappa_T}{\omega R^2}. \quad (20)$$

When the mass transfer is controlled by the diffusion of the volatile component in the liquid phase, the diffusion

equation in the liquid and the equilibrium condition at the bubble surface are added. Then the effective bulk modulus which includes both the heat and mass transport is

$$\frac{p_{go}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma} i\chi \left[ \sqrt{\frac{i}{\chi}} \coth \left( \sqrt{\frac{i}{\chi}} \right) - 1 \right] - 3A_g \sqrt{\alpha_g} i\chi \left( \sqrt{\frac{i}{\chi}} + \sqrt{\alpha_g} \right), \quad (21)$$

$$A_g = \frac{\rho_l p_{go}}{\rho_{go}} \frac{\partial C_{eq}}{\partial p}, \quad (22)$$

$$\alpha_g = \frac{\kappa_{gl}}{\kappa_T}, \quad (23)$$

where  $C_{eq}$  is the saturation concentration at  $p_{go}$  [30]. Equation (21) has the last term in addition to Eq. (19), which represents the effect of the mass transfer. The dimensionless parameter,  $A_g$ , represents the ratio of the volatile mass going into the gas phase from a unit volume of the liquid phase by decompression to the mass change in a unit volume of the gas phase due to expansion.

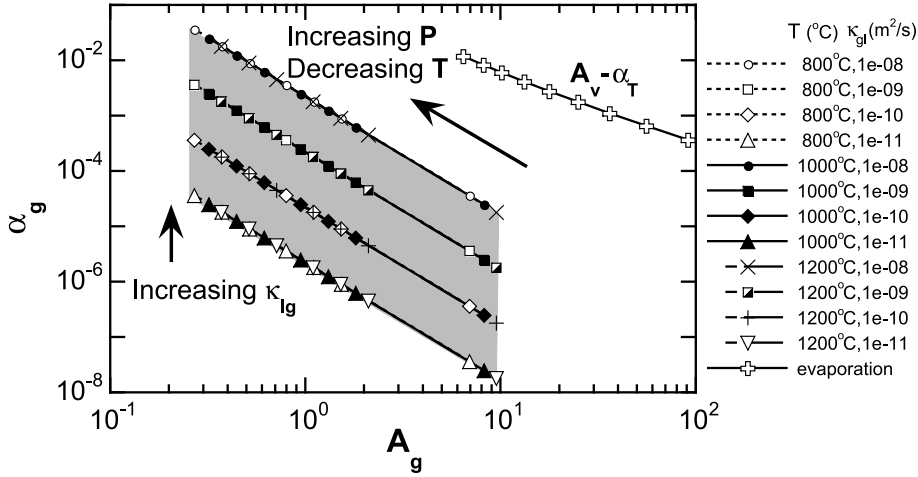
Figure 2 shows the relevant range of the dimensionless parameters,  $A_g$  and  $\alpha_g$ , for an  $\text{H}_2\text{O}$  bubble in magma [30]. As temperature decreases or pressure increases,  $A_g$  decreases (Fig. 2) because of the following two reasons. With decreasing temperature,  $\rho_{go}^{-1}$  decreases. In the case of magma,  $C_{eq}(p)$  is approximately proportional to  $\sqrt{p}$  [26] so that  $\partial C_{eq}/\partial p$  decreases with increasing pressure.

The effective bulk modulus of a bubble for some selected values of the parameters in the range is presented in Fig. 3 [30]. The thick broken lines in the figure are obtained by Eq. (19), which includes only the heat transport. In this case, the real part approaches the isothermal bulk modulus and the adiabatic one in the low and high frequencies, respectively. The mass transport makes the bubble stiffness ( $\text{Re}(K_g)$ ) smaller, which is the more significant in the lower frequency regime. It is because the bubble has more time to take in and out the volatile from the liquid in a cycle of the pressure perturbation. The imaginary part for each parameter set has a local peak around

$$\omega \sim \tau_T^{-1} = 15\kappa_T R_o^{-2}, \quad (24)$$

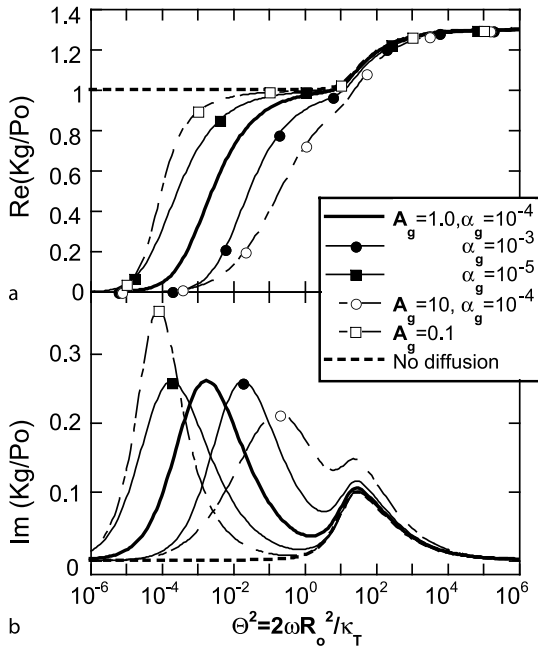
which is the characteristic frequency of the energy loss due to heat transfer. In the case of  $R_o = 10^{-3} \text{ (m)}$  and  $\kappa_T = 4 \times 10^{-6} \text{ (m}^2 \text{ s}^{-1}\text{)}$ , which is the value for  $\text{H}_2\text{O}$  at 10 MPa and 1273 K [5], the corresponding frequency ( $\omega/(2\pi)$ ) is 9.5 Hz. The imaginary part of  $K_g$  including the diffusion effect has another peak at the characteristic frequency of the energy loss due to the mass transport. The frequency is approximately represented by

$$\omega \sim \tau_g^{-1} = 9\alpha_g \kappa_T A_g^2 R_o^{-2}, \quad (25)$$



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 2**

The relevant range of the dimensionless parameters representing the effect of the volatile transfer in the magmatic system is shown as the *gray area*. The parameter  $A_g$  and  $\alpha_g$  are defined in Eqs. (22) and (23), respectively. The temperature ( $T$ ) and the volatile diffusivity ( $\kappa_{gl}$ ) are assumed as shown in the legend, and the pressure is varied from 0.1 to 100 MPa. The *open crosses* are the corresponding parameters for a vapor-bubble system,  $A_v$  and  $\alpha_T$  given in Eqs. (27) and (28), respectively. The temperature is varied from 380 K to 500 K and the pressure is the saturation pressure at each temperature. (Modified from Fig. 1 in [30])



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 3**

The effective bulk modulus of a gas bubble with heat and mass diffusion (calculated by Eq. (21)) as a function of the dimensionless frequency. The real and the imaginary parts are presented in **a** and **b**, respectively. The *thick broken lines* indicate no diffusion and include only thermal effects (which is calculated by Eq. (19)). (Fig. 2 in [30])

which is usually smaller than  $\tau_T^{-1}$  [30]. It is noted that the above model assumes a single bubble in an infinite liquid. When the oscillation period is very long, interaction of the diffusion layers of the adjacent bubbles has to be considered [15].

When the mechanism of the mass exchange between the liquid and the bubble is the evaporation/condensation at the bubble wall, the latent heat plays an important role. Then the thermal diffusion equation in the liquid and the balance between the heat flux through the bubble surface and generation of the latent heat should be added [19,24]. The corresponding bulk modulus of the bubble is represented as [24]:

$$\frac{p_{go}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma} \left(1 - \frac{c_{pg}T_o}{L}\right)^2 i\chi \left[ \sqrt{\frac{i}{\chi}} \coth \left( \sqrt{\frac{i}{\chi}} \right) - 1 \right] - 3A_v \sqrt{\alpha_T} i\chi \left( \sqrt{\frac{i}{\chi}} + \sqrt{\alpha_T} \right), \quad (26)$$

$$A_v = \frac{\rho_l c_{pl} T_o p_{go}}{(\rho_{go} L)^2}, \quad (27)$$

$$\alpha_T = \frac{\kappa_{Tl}}{\kappa_T}, \quad (28)$$

where  $c_{pg}$  and  $c_{pl}$  are the heat capacity at constant pressure in the gas and the liquid phases, respectively,  $L$  is the latent heat, and  $\kappa_{Tl}$  is the thermal diffusivity in the liquid. In obtaining Eq. (26), the Clausius-Clapeyron re-

lation:  $(dp/dT)_{\text{sat}} = L\rho_g/T$ , and thermodynamic relations for an ideal gas are used.

Equation (26) has the same form as Eq. (21) except  $c_{pg}T_o/L$ . This difference is due to the temperature change at the bubble wall. The dimensionless parameter,  $A_v$ , corresponds to  $A_g$  and has a similar physical meaning, which represents the ratio of the mass going through a phase change in a unit volume of the liquid phase to the mass change in a unit volume of the gas phase due to expansion. Although the equation is similar, the possible range of the parameter is different (Fig. 2). As a result, the frequency dependence of  $K_g$  is also different as is shown in Fig. 4. Comparing the figure with Fig. 3, we can see that the energy loss of the vapor bubble due to the phase change is significant in a frequency range higher than that due to diffusion. The frequency range is comparable to that of the heat transfer, but the amount of energy loss is much larger. The vapor bubble loses its elasticity, which is represented by  $\text{Re}(K_g)$ , in the lower frequency as well.

Under the action of the sound field, there is a net transport of heat into the bubble by a non-linear pro-

cess called the rectified heat transfer [98]. In the evaporation/condensation system, the order of the non-linear effect is so large that it affects the amplitude and damping of the oscillation in the linear regime [24]. Equation (26) does not include the effect. Some works investigating the effect of rectified diffusion process in triggering an eruption or an earthquake are introduced in the later sections.

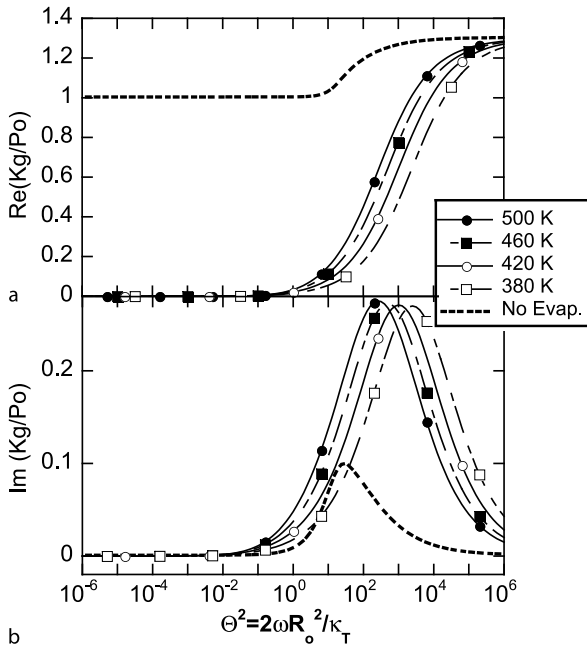
### Translational Motion

So far we have neglected the translational motion of a bubble relative to the liquid. The translational motion is considered to be negligible when the translational displacement is smaller than the diffusion layer in the liquid surrounding the bubble. In an acoustic field with frequency  $\omega$ , this condition is represented by  $U/\omega < \sqrt{\kappa/\omega}$ , where  $U$  is the translational speed and  $\kappa$  is the relevant diffusivity. The translational velocity of a spherical bubble driven by buoyancy is estimated by  $U = k\rho_l R_o^2 g \eta_l^{-1}$ , where  $g$  is the gravitational acceleration. Although  $k = 1/3$  for a pure liquid,  $k = 2/9$  is used for most of actual liquid, which is not perfectly pure, because the pro-surface components concentrate on the bubble surface to make the surface less mobile [49]. These approximations hold for relatively slow velocity, which satisfies  $Re_t = 2\rho_l R_o U \eta_l^{-1} \leq 1$ . Then the condition in which the translational motion has a minor effect on the heat and mass transfer is

$$\omega > \frac{R^4}{\kappa} \left( \frac{k\rho_l g}{\eta_l} \right)^2. \quad (29)$$

Assuming  $k = 2/9$ ,  $R = 10^{-3}$  (m),  $\kappa = \kappa_{gl} = 10^{-11}$  ( $\text{m}^2 \text{s}^{-1}$ ),  $\rho_l = 2500$  ( $\text{kg m}^{-3}$ ),  $\eta_l = 10^5$  (Pa s) for a magma- $\text{H}_2\text{O}$  system,  $\omega > 3 \times 10^{-4}$  ( $\text{rad s}^{-1}$ ) and  $U = 6 \times 10^{-6}$  ( $\text{m s}^{-1}$ ). For a water-vapor system, on the other hand, we assume  $k = 1/3$ ,  $\kappa = \kappa_{tl} = 10^{-7}$ ,  $\rho_l = 1000$ ,  $\eta_l = 10^{-3}$ . Then, if  $R = 10^{-5}$ ,  $\omega > 1$  and  $U = 3 \times 10^{-4}$ , and if  $R = 10^{-4}$ ,  $\omega > 10^4$  and  $U = 3 \times 10^{-2}$ . According to these estimations, we can see that the translational motion is negligible for most cases with magma except basalt, which has relatively small viscosity ( $\eta_l < 10^2$ ) and large diffusivity ( $\kappa_{gl} \sim 10^{-9}$ ), while it is considerable in hydrothermal systems, except for very small bubbles and the time scale is very short. As an example, the effect on the thermal collapse of a vapor bubble is introduced later.

Another effect of the translational motion is its mechanical coupling with the radial motion. Because a bubble has to move the surrounding liquid in order to make itself move, it is subject to the inertial force of the liquid, which depends on its volume [49]. Therefore when the bubble volume changes, the force also changes. By this consideration, an equation of the translational motion of the bubble



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 4**

The effective bulk modulus of a vapor bubble with thermal and evaporation effects (calculated by Eq. (26)) as a function of the dimensionless frequency. The real and the imaginary parts are presented in a and b, respectively. The thick broken lines indicate no evaporation and include only thermal effects (which is calculated by Eq. (19))

is approximately represented as [101]:

$$\dot{U} = -\frac{3}{R}\dot{R}U + 2g - \frac{3}{4}\frac{C_D}{R}|U|U, \quad (30)$$

where  $C_D$  is the non-dimensional drag coefficient, which is given as a function of  $Re_t$ . From the first term in the right-hand side, we see that the translational motion is decelerated by the bubble expansion. Although it had been theoretically recognized for long time, it was quantitatively justified by experiments recently [66]. On the other hand, the effect of the translational motion on the radial motion is represented by the term,  $\rho U^2/2$ . This term can usually be neglected [37,66] except in cases with very strong oscillation [17].

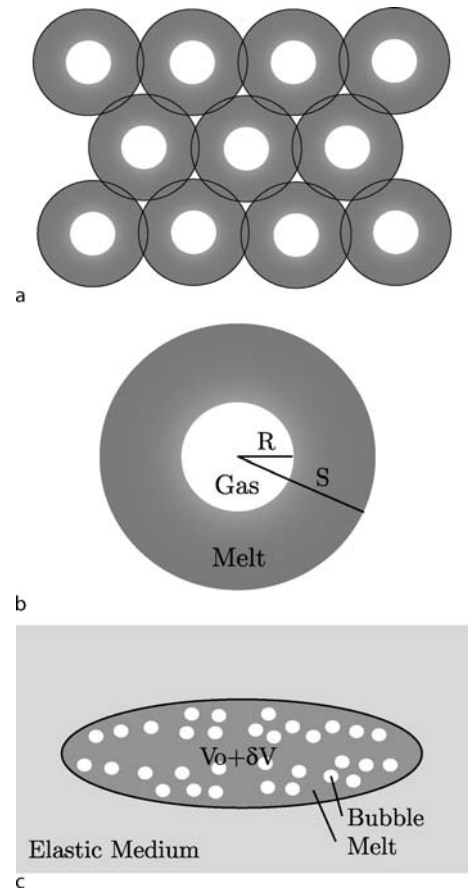
### Bubbly Magma in an Elastic Rock as a Pressure Source

#### Model Overview

Here we consider a magma chamber filled with compressible viscous melt and numerous tiny  $H_2O$  gas bubbles (Fig. 5) [79,87]. The magma chamber is confined in an elastic rock. When perturbation is given to the system, pressure may increase by interaction of the elastic deformation of the chamber, expansion of the bubbles, and gas diffusion from the melt to the bubble. Recently, the process has been discussed in the literature in relation to the observed volcanic phenomena [6,13,29,53,64,87,97], which are presented in the introduction.

The melt and bubbles are expressed by the cell model [79], in which multiple spherical bubbles of a constant radius are uniformly packed. Each bubble is surrounded by a finite volume of the melt, represented by an elementary cell. The elementary cell is spherical, in which a single gas bubble is located at the center. It is assumed there is no interaction between neighboring elementary cells such that all gas bubbles grow in the same manner. This simplification enables us to examine bubble growth processes in the entire chamber by studying the growth of just a single bubble, which is represented by Eq. (5).

The main mechanism for increasing the pressure is diffusion of the volatile. It is the slowest process of the bubble dynamics as is described in the previous section. It is certainly longer than the period of resonant oscillation of the individual bubbles so that the inertia terms in Eq. (5) are neglected. It is also longer than the time scale of the heat transport within the bubble as is shown in Fig. 3 so that we may assume uniform and constant temperature within the bubble. Then the mathematical model for the elementary cell consists of three equations, which represent the radial motion of the bubble, volatile diffusion in the melt, and



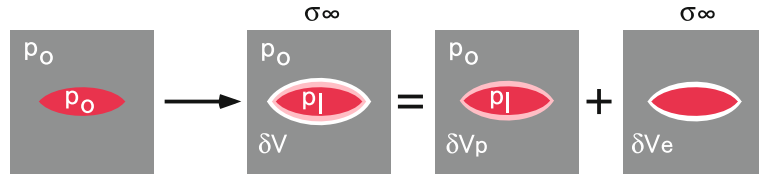
**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 5**

Schematic illustrations of a cell model [79], **b** an elementary cell, and **c** a chamber surrounded by an elastic medium. The chamber is filled with compressible viscous melt and numerous tiny spherical gas bubbles. Magma is represented by a combination of many elementary cells.  $R$  and  $S$  is the radius of the elementary cell and gas bubble, respectively, and  $V_0 + \delta V$  is the volume of the chamber. (Modified from Fig. 1 in [87])

ideal gas approximation, respectively, and three boundary conditions, which are phase equilibrium and mass flux at the bubble surface and no mass flux at the external boundary of the cell element.

#### Interaction Between Melt and Elastic Medium

The volumetric change of the bubbles and the melt due to the pressure change is compensated by the elastic deformation of the chamber. Here we consider the initial pressure and stress conditions in relation to the physical process which brings about the condition, since the relations have not always been mentioned clearly in previous litera-



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 6**

**Mathematical representation for the interaction of melt pressure  $p_l$ , stress in the ambient rock  $\sigma_\infty$ , and volume change of the chamber,  $\delta V$ . Volume change due to internal overpressure  $p_l - p_o$  and that due to external stress is considered separately**

ture. We assume quasi-static deformation of the chamber, where the pressure of the melt is balanced by the elastic stress applied by the wall of the chamber. The volumetric change can be caused by (a) pressure change within the chamber and (b) stress change in the surrounding rock (Fig. 6). Each process is individually represented by

$$p_l - p_o = \bar{\mu}_p \delta V_p / V_o, \quad (31)$$

$$-\sigma_\infty = \bar{\mu}_e \delta V_e / V_o, \quad (32)$$

where  $V_o$  is the initial equilibrium volume of the chamber,  $\sigma_\infty$  is the ambient stress change,  $\delta V_p$  and  $\delta V_e$  are the volumetric change due to (a) and (b), respectively, and  $\bar{\mu}_p$  and  $\bar{\mu}_e$  are the corresponding effective stiffness of the wall. Each effective stiffness depends on the elasticity of the rock, the geometry of the chamber, and the applied stress field. For the simplicity, we approximate  $\bar{\mu}_p = \bar{\mu}_e = \bar{\mu}$ . Then the total volumetric change,  $\delta V$ , is given by

$$p_l - p_o - \sigma_\infty = \bar{\mu} \delta V / V_o. \quad (33)$$

The two perturbations which cause the volumetric change have not been clearly distinguished in previous literature. The mathematical treatment by [87] assumed that  $p_l - p_o = -\Delta p$  is given at  $t = 0$ . They considered that this pressure drop is caused by a decrease of the ambient stress field by a certain amount, say  $\sigma_\infty = -\Delta\sigma$ . On the other hand, the assumption of [13] is that the pressures in all of the bubbles, the melt, and the rock are lower by  $\Delta p$  than the saturation pressure for the dissolved volatile concentration at  $t = 0$ . Strictly speaking, the consequent processes are different depending on what causes the pressure perturbation. If the pressure drop of  $\Delta p$  occurred first within the chamber, that is  $p_l - p_o = -\Delta p$  and  $\sigma_\infty = 0$ , the chamber would initially shrink according to Eq. (33). If it is caused by an ambient stress drop first, that is  $p_l - p_o = 0$  and  $\sigma_\infty < 0$ . Then the chamber would initially expand. In either case, the initial response is almost instantaneous, which is controlled by elasticity of the rock and compressibility of the melt. The major deformation

occurs later and is controlled by volumetric change of the bubbles.

### Response to Sudden Decompression and Characteristic Time for Pressure Recovery

There are three important parameters to characterize the response of the system to the pressure drop which are useful for comparing the model and the field observations. The first one is the re-equilibrated pressure  $p_f$ . The second is the final bubble radius,  $R_f$ . The third is the characteristic time of the recovery process,  $T_{\text{growth}}$ .

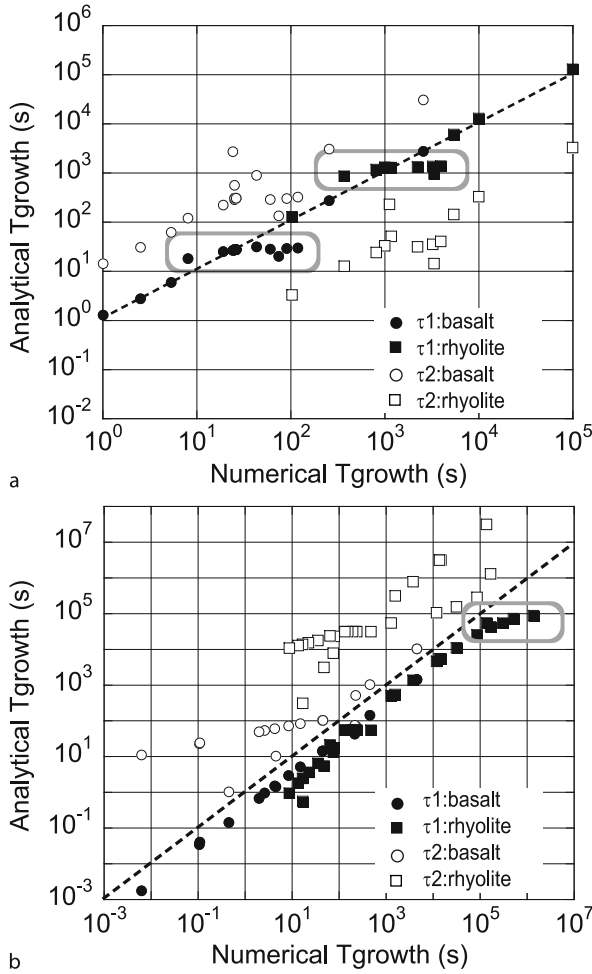
The first and the second are calculated by consideration of the equilibrium condition alone and can be calculated semi-analytically [64]. On the other hand,  $T_{\text{growth}}$  is determined by numerical calculation of the set of equations described above. Shimomura et al. [87] investigated the recovery processes and presented that  $T_{\text{growth}}$  depends on the stiffness of the chamber ( $\bar{\mu}$ ), initial bubble radius ( $R_o$ ), number density of the bubbles ( $N$ ), volatile diffusivity in the melt ( $\kappa_{\text{gl}}$ ), initial pressure ( $p_o$ ), the pressure drop ( $\Delta p$ ), and the melt properties in a complicated manner.

The corresponding study for the bubble growth in an open space, where the pressure is constant regardless of the bubble expansion, was presented by Prousevitch et al. [79]. They assumed an initially supersaturated condition, in which both  $p_{l0}$  and  $p_{go}$  are lower than the saturation pressure of the volatile dissolved in the melt. They investigated the final bubble radius and the time to reach it, which correspond to  $R_f$  and  $T_{\text{growth}}$ , respectively. They also presented effects of initial bubble radius ( $R_o$ ), number density of the bubbles ( $N$ ), volatile diffusivity in the melt ( $\kappa_{\text{gl}}$ ), initial pressure ( $p_o$ ), and initial super-saturation.

A simple theory to estimate the time scale of re-equilibration is useful to compare the model with observations, but has not been determined yet. Here we test two hypotheses.

1. The recovery time is comparable with the time scale in which the diffusion layer develops over the entire shell, that is  $\tau_1 = (S_f - R_f)^2 / \kappa_{\text{gl}}$ .





**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 7**

Numerical results of the pressure recovery time for magma in an elastic rock **a** and the bubble growth time in an open system **b** for various system parameters are compared with analytical approximations:  $\tau_1$  is the time scale of mass diffusion across the final shell thickness,  $\tau_2$  is approximation by [53]. Agreement between the numerical results and  $\tau_1$  is better, but some systematic discrepancy remains, as indicated by gray frames. The numerical results for **a** are from [87], and those for **b** are from [79]

- Based on a dimensional analysis of the simplified diffusion Eq. (14), Lensky et al. [53] proposed  $\tau_2 = (R_f^2/\kappa_{gl})(\rho_{gf}/\rho_l)(C_o - C_f)^{-1}$ , where  $\rho_{gf}$  and  $C_f$  are the final gas density in the bubble and the volatile concentration remained in the melt, both of which are functions of  $p_f$ . They obtained this equation based on the approximation that the quasi-static mass flux through the interface is  $(C_o - C_{eq}(p_g))/R$ .

The re-equilibration times,  $T_{\text{growth}}$ , obtained by Shimomura et al. [87] and Prousevitch et al. [79] are compared

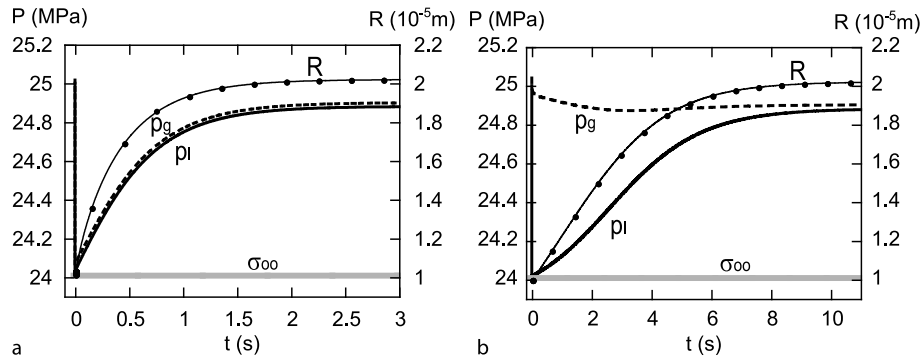
with the above models in Fig. 7. Comparing  $T_{\text{growth}} - \tau_1$  (black symbols) and  $T_{\text{growth}} - \tau_2$  (white symbols), we see that the general trend of  $T_{\text{growth}}$  is better estimated by  $\tau_1$  than by  $\tau_2$  in both confined and open systems. However, it should also be noted that  $\tau_1$  still has systematic errors which are indicated by gray frames. The errors are more dominant in the confined system (Fig. 7a). It is indicated that the simple estimation does not include all the factors relevant to the re-equilibration time and it is not necessarily applicable to the wider range of the parameters.

### Re-equilibration Processes

Here we discuss different re-equilibration processes depending on the cause of the pressure drop and the relevant initial conditions. Three representative solutions are presented in Figs. 8–10. They are obtained for the standard basaltic system [87], but only viscosity is varied from 50 Pa s for (a) to  $10^6$  Pa s for (b).

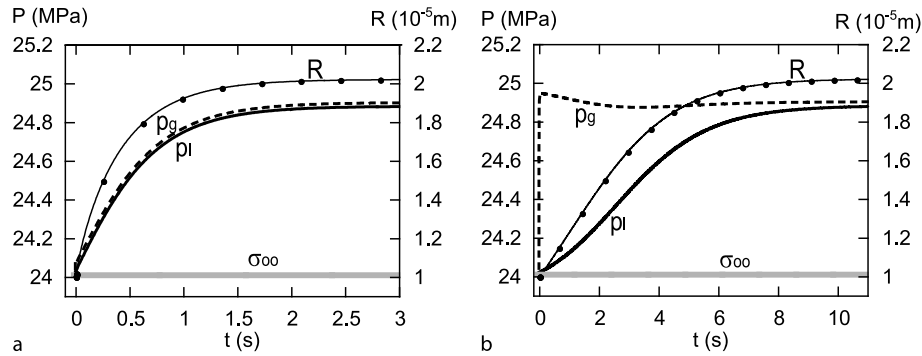
Figure 8 is the case in which the stress drop occurs in the ambient rock first. It is generated by, for example, surface unloading by dome collapse [97] and stress change after a local earthquake. The condition is represented by  $\sigma_\infty = -\Delta\sigma$  at  $t \geq 0$  while  $p_l = p_g - 2\Sigma/R_o = p_o$  at  $t = 0$ . According to Eq. (33), the chamber expands instantaneously, and  $p_l$  drops by  $\Delta p$ . Then the initial condition assumed by [64,87] is attained. Response of  $p_g$  is not instantaneous [62]. Due to the difference between  $p_g$  and  $p_l$ , the bubble expands according to Eq. (5) to decrease  $p_g$ . Then the difference between  $p_g$  and the equilibrium pressure for the volatile concentration in the melt occurs to make the volatile flow into the bubble to re-increase  $p_g$ . As the bubbles grow, the entire volume of the magma ( $\delta V$ ) increases to enlarge the chamber elastically. Then the elastic stress  $\bar{\mu}\delta V/V_o$  increases  $p_l$  according to Eq. (33). The re-equilibration proceeds in this way [87].

Figure 9 is the case in which the pressure in the bubble as well as those in the melt and the ambient rock is lower than the saturation pressure for the initial volatile concentration in the melt at  $t = 0$ . This condition occurs if bubbles are mixed with the supersaturated melt instantaneously, or if the bubbles are kept in the supersaturated mixture without interaction and suddenly allows the diffusion. Mathematically, the initial condition is equivalent to those assumed by [13] and [79]. The condition is represented by  $p_l - p_o = p_g - 2\Sigma/R_o - p_o = \sigma_\infty = -\Delta p$  at  $t \geq 0$ . Diffusion of the volatile into the bubble starts, which increases  $p_g$  first. Then  $p_g - p_l$  expands the bubble and the chamber to increase  $p_l$  in the same way as the previous case. Practically, the difference between Fig. 8 and Fig. 9 occurs only during a very short period in the beginning,



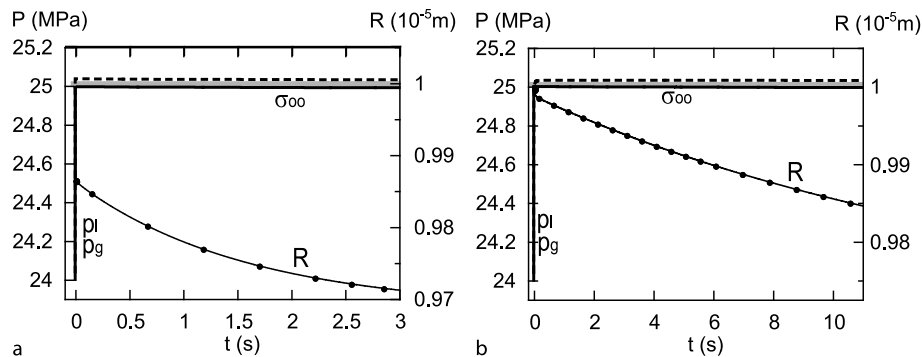
**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 8**

Pressure recovery in a bubbly magma in an elastic chamber after sudden unloading  $\sigma_{\infty} = -1$  MPa. The initial condition is  $p_l = p_g - 2\Sigma/R = 25$  MPa and  $R = 10^{-5}$  m. The bubble radius on the right axis is plotted with a line and points. The stress and pressures on the left axis are plotted with a solid line for  $p_l$ , a dotted line for  $p_g$ , and gray line for  $\sigma_{\infty}$ . The system parameters are  $\kappa_{gl} = 10^{-8}$  m<sup>2</sup> s<sup>-1</sup>, the bubble number density is  $10^{11}$  m<sup>-3</sup>, and  $\eta_l = 50$  Pa s for a and  $10^6$  Pa s for b. The others are the same as those for the basaltic system by [87]



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 9**

Similar to Fig. 8, but the initial condition is  $\sigma_{\infty} = p_l - p_o = p_g - 2\Sigma/R - p_o = -1$  MPa, with  $p_o = 25$  MPa



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 10**

Similar to Fig. 8, but the initial condition is  $\sigma_{\infty} = 0$  and  $p_l - p_o = p_g - 2\Sigma/R - p_o = -1$  MPa, with  $p_o = 25$  MPa

and the subsequent increase of the pressure and volume of the chamber may look the same from outside.

Figure 10 is the case in which the pressure drop occurs in the melt and in the bubble, while the stress in the ambient rock is unchanged. The condition is represented

by  $p_l - p_o = p_g - 2\Sigma/R_o - p_o = -\Delta p$  and  $\sigma_{\infty} = 0$ . Although the assumed initial condition is rather imaginary, this case is presented in order to demonstrate how the response can be different depending on the way the system is decompressed. In fact, it is more realistic that  $p_l$  drops

first, while  $p_g$  remains at the initial value. This situation may occur by small leakage of the melt from the system. In this case, the melt pressure just recovers almost instantaneously, because the container compresses the melt according to Eq. (33) and bubbles also compress the melt. No other significant change is expected. On the other hand, if both  $p_l$  and  $p_g$  drop, as in Fig. 10, the pressure still recovers rapidly, but bubbles are compressed. Because the mechanical balance of the bubble and the melt is attained with  $p_g - 2\sigma/R = p_l$  according to Eq. (5),  $p_g$  has to become larger when  $R$  decreases. Then  $p_g$  exceeds the equilibrium pressure for the volatile in the melt to make the volatile dissolve into the melt.

### Rectified Diffusion

So far, we discussed responses of the system to a stepwise pressure drop. When the perturbation is caused by a seismic wave from an external source, the system is subject to a cyclic disturbance. Rectified diffusion is a mechanism which can push dissolved volatiles into bubbles in a sound field. Bubbles take in more volatiles during expansion than they discharge during contraction, mainly because of the following two non-linear effects [18,27]. Firstly, the interface is larger during expansion than during contraction. Secondly, radial bubble expansion tangentially stretches the diffusion layer and sharpens the radial gradient of the volatile concentration in the diffusion layer, so that the volatile flux into the bubble is enhanced.

Brodsky et al. [6] discussed the possible pressure increase of a bubbly magma confined in an elastic rock by this mechanism. Using the solution by Hsieh and Plesset [27] for a periodic system, they considered that, even though the net pressure changes are determined by the pre-existing oversaturation, the rectified diffusion accelerates the pressure increase and may break the balance which had been stabilized the system prior to the oscillation. Ichihara and Brodsky [29] improved the solution by including resorption of gas as the pressure increase and development of the diffusion layer around the bubble in a self-consistent way. It is then shown that rectified diffusion is not faster than the ordinary diffusion and its contribution to the net pressure change is at the most  $2 \times 10^{-9}$  of the initial pressure regardless of the pre-existing oversaturation.

## Acoustic Bubbles in Hydrothermal Systems

### Pressure Impulses Generated in a Geyser

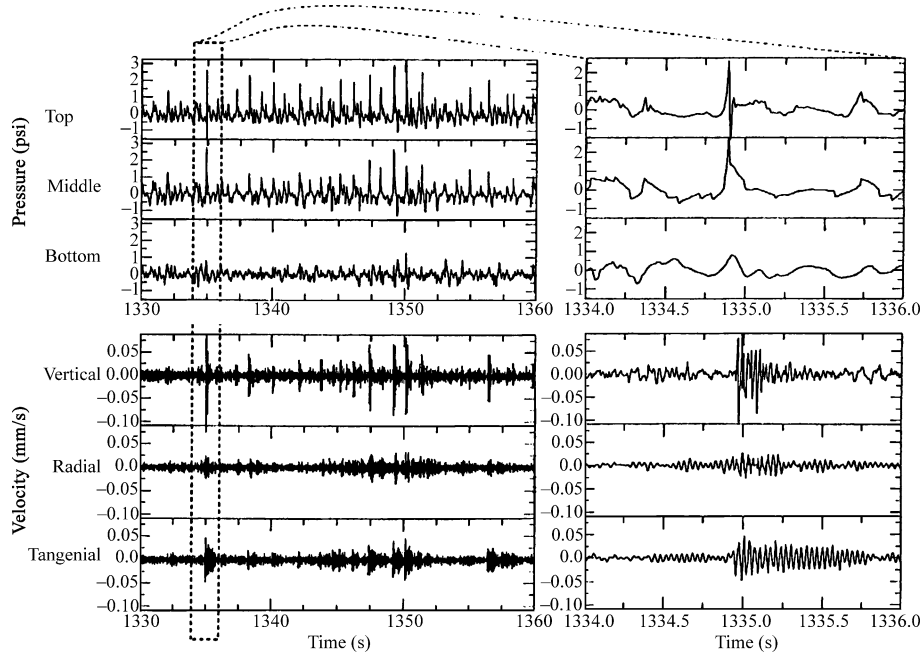
Here we consider a mixture of water and vapor bubbles, in which we expect effects of bubble oscillations and evaporation.

Kedar et al. [39,40] conducted field experiments at Old Faithful Geyser, Yellowstone. They measured pressure within the geyser's water column simultaneously with seismic measurements on the surface. The data show a distinct cause-and-effect relationship between the impulsive pressure source and the impulse response of the rock surrounding the water column. In addition, the pressure pulse, which is strongest at the top transducer, strongly attenuates downward. Considering that the pulse is generated by the oscillation of a single bubble, they compared one selected signal with a solution of the equation of motion for the bubble radius (Eq. (3)). In order to fit the measured oscillation with a reasonable bubble radius, they had to assume a very small ambient pressure to lower the frequency, and a very large viscosity to increase the damping. For  $R_o = 0.055$  m, for example, they used  $p_{go} = 0.02$  MPa, with which Eq. (9) gives the resonant frequency close to the observation:  $\sim 20$  Hz. The viscosity was assumed as  $\eta_l = 40$  Pa s, which is larger than the actual value by more than four orders. They compared the damping coefficient with those from radiation and heat transfer, though these effects were not included in the calculation of Eq. (3), and concluded that mechanisms other than acoustic, thermal, or viscous damping are required to explain the strong damping observed.

We have already introduced the damping coefficient  $b_t$  with the evaporation effect in Eq. (17) with Eq. (26). Then, assuming the similar bubble radius and frequency ranges as [40], let us see the damping coefficient by evaporation in comparison with the other coefficients, which are for viscous, acoustic and thermal damping, represented by Eqs. (8), (13), and (17) with Eq. (19), respectively. Their values are compared in Fig. 12b, assuming  $p_{go} = 0.13$  MPa (the saturation pressure at 380 K). We can see that the evaporation effect significantly increases the damping and dominates the other damping mechanisms in the frequency range of the geyser oscillation. The evaporation effect also decreases  $\text{Re}(K_g)$  (Fig. 12a) in the range. It is thus suggested that the evaporation effect is significant for the bubble dynamics in the hydrothermal system.

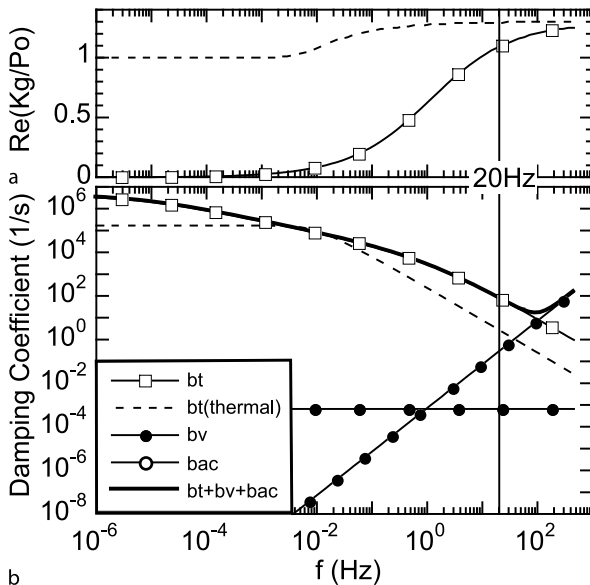
### Inertial and Thermal Collapse of a Bubble

When we heat water in a kettle, we hear strong intermittent pulses before boiling starts. The phenomenon is explained in terms of the bubble dynamics as follows [1]. In the first regime (which initiates above approximately 40 °C), small bubbles form slowly out of dissolved air in the liquid, rising silently to the surface as they break off the side of the vessel. At higher temperature ( $\sim 70$  °C), vapor



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 11**

Simultaneous pressure records (through a high-pass filter at 1 Hz) and seismic traces at Old Faithful geyser, Yellowstone. The geyser's eruptions are 2–5 min long with the interval between them ranging from 30–100 min. The figure shows a 30 s data about 27 min after the previous eruption and about 52 min before the next eruption. The conduit of the geyser is 22 m deep, where the bottom sensor was located. The bottom, middle and top sensors were connected 3 m apart. The seismic station was located at  $\sim 25$  m from the geyser. The data show a direct correspondence between the pressure pulses and the seismic signals that follow them. (Fig. 3 in [39])

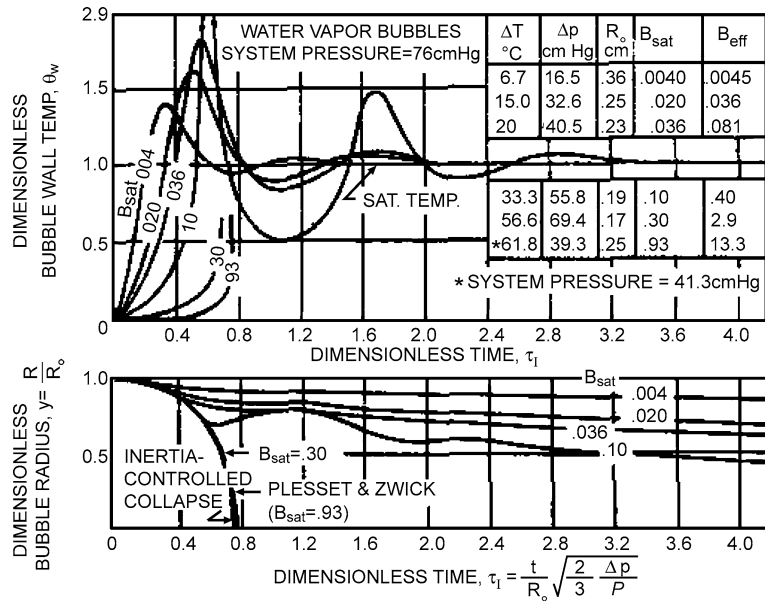


**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 12**

The bubble elasticity (a) and damping factors (b) for a single vapor bubble with radius 0.055 m at 380 K, 0.13 MPa (saturation pressure)

bubbles start to nucleate at various sites at the heated bottom surface of the container. Vapor bubbles are different from the air bubbles in that their formation and collapse (at the bottom of the vessel) occurs explosively, producing pressure impulses that traverse the liquid and cause much of the sound we hear. In the third stage (between 90 °C and 100 °C), vapor bubbles grow, coalesce, and survive their ascent through the liquid. Bursting of vapor bubbles at the top surface is considered to be the sound source in this regime. Finally, the transition to full boil is characterized by large bubble formation throughout the bulk of the liquid.

We consider whether and how the impulse generation by the bubble collapse occurs in a geyser, where water that is already boiling is injected and cooled from the surface [43]. The collapse (or growth) of a bubble is classified into two modes: the inertia mode, which is controlled by the liquid inertia and driven by the pressure difference between the liquid and the bubble, and the thermal mode, which is controlled by the heat transfer and driven by the temperature difference [20,103]. The former is more violent than the latter and is responsible for the impulse generation.



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 13**

Variation of wall temperature and radius during collapse of water vapor bubbles.  $B_{sat}$  is the dimensionless parameter given by Eq. (35), which determines relative importance of the liquid inertia and the heat transfer (Fig. 3 in [20])

Based on theoretical and experimental studies, Florschuetz and Chao [20] proposed that the relative importance of the inertia and the heat transfer is evaluated by a dimensionless parameter,  $B$ , defined by

$$Ja = \frac{\rho_l c_{pl}(T_{sat}(p_{lo}) - T_o)}{\rho_{go} L}, \quad (34)$$

$$B = Ja^2 \frac{K_{TI}}{R_o} \sqrt{\frac{\rho_l}{p_{lo} - p_{sat}(T_o)}}, \quad (35)$$

where  $T_{sat}(p_{lo})$  is the saturation temperature for the ambient pressure ( $p_{lo}$ ) and  $p_{sat}(T_o)$  is the saturation pressure for the system temperature ( $T_o$ ). The dimensionless parameter  $Ja$  is called the Jacob number, which represents the degree of subcooling. Figure 13 displays their calculation results, which clearly shows that for  $B \geq 0.3$ , the collapse rate is dominated by liquid inertia effect, while for  $B \leq 0.03$  it is much slower and is recognized as the thermal mode. For an intermediate value of  $B$ , oscillation is observed.

In these works [20,77], it is often assumed that the pressure in the bubble is initially equal to the saturation pressure of the subcooled liquid, that is  $p_{go} = p_{sat}(T_o)$ , which is less than the ambient pressure ( $p_{lo}$ ) [20,77]. Experimentally, it is achieved by preparing for a thermally equilibrium water-vapor system at a low pressure and suddenly increasing the system pressure to  $p_{lo}$  [20]. Then the initial collapse is relatively violent and continues by inertia until the vapor heating at the bubble wall increases the

vapor pressure above the ambient pressure to such an extent that the liquid is momentarily brought to rest and its motion actually reverses before the vapor pressure again drops below the system pressure [20,77]. Although the oscillation is difficult to see on the radius change curves in Fig. 13 for small  $B$ , the beginning inertia controlled stage is evidenced by that all the curves start along the inertia curve.

On the other hand, in case that a bubble suddenly enters cold water,  $p_{go} = p_{lo} > p_{sat}(T_o)$  while temperature in the bubble  $T_{go}$  is larger than  $T_o$  and is close to  $T_{sat}(p_{go})$ . Then the collapse begins in the gentle mode controlled by the heat transfer. It can turn into the inertia mode only if the rate of heat transport and condensation to decrease the vapor pressure is so large that inward motion of the surrounding liquid cannot follow. Prosperetti and Hao [77] presented that relative translational motion between the bubble and the liquid significantly increases the rate of heat transport and accelerates the bubble collapse. Furthermore, they pointed out the coupling effect between the translational and the radial motions. As Eq. (30) suggests, the decreasing bubble radius ( $\dot{R} < 0$ ) works to accelerate the translational motion. Although the drag force ( $\propto C_D R^{-1} |U|U$ ) increases as  $R$  decreases and  $U$  increases, there are cases in which the contribution of the first term is so large as to make  $\dot{U} > 0$ . Then the collapse and the translational motion of the bubble accelerate each other [77].



### Rectified Heat Transfer

In the same way as the rectified diffusion discussed in the previous section, rectified heat transfer works for a vapor bubble in an acoustic field [24,98]. When the bubble is compressed, some vapor condenses, the surface temperature rises, and heat is conducted away into the adjacent liquid. When the bubble expands during the following half cycle, evaporation causes a temperature drop of the bubble surface, with a consequent heat flux from the liquid. The imbalance of the heat flux and the interface area between the compression phase and the expansion phase causes the net energy flux into the bubble.

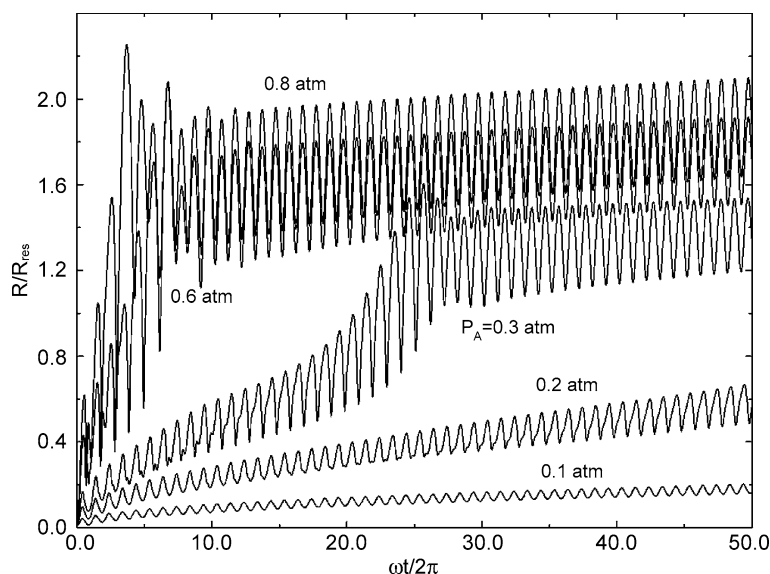
Sturtevant et al. [89] investigated the effect of rectified diffusion on pressure increase in a hydrothermal system, as a possible mechanism for triggered seismicity by a distant earthquake. They modeled the system as a two-component  $\text{H}_2\text{O}-\text{CO}_2$  system, and considered rectified mass diffusion. As is mentioned in the previous section, the net pressure change due to rectified mass diffusion is very small, if it is evaluated in a self-consistent way [29].

Although rectified heat transfer has a similar mechanism as the rectified mass transfer, it is much more intense since the thermal diffusivity of liquids typically exceeds the mass diffusivity by two orders of magnitude [77]. It can grow a vapor bubble quickly within several cycles of oscillation at the beginning (Fig. 14), and thus can be effective even with low-frequency pressure waves. Moreover the effect is reinforced by the coupling effect of the bub-

ble growth and translational motion [25], and coupling of evaporation and diffusion of another gas component [38]. The net energy flux into the bubble increases the temperature in the bubble, which may change the liquid static pressure [24]. It is noted that numerical results in the literature cannot be directly used to estimate the pressure increase, because they were obtained for an open system in which the bubble continues to grow instead of increasing system pressure. However, it might be worth while to re-evaluate effects of rectified processes in a hydrothermal system taking account of these recent results.

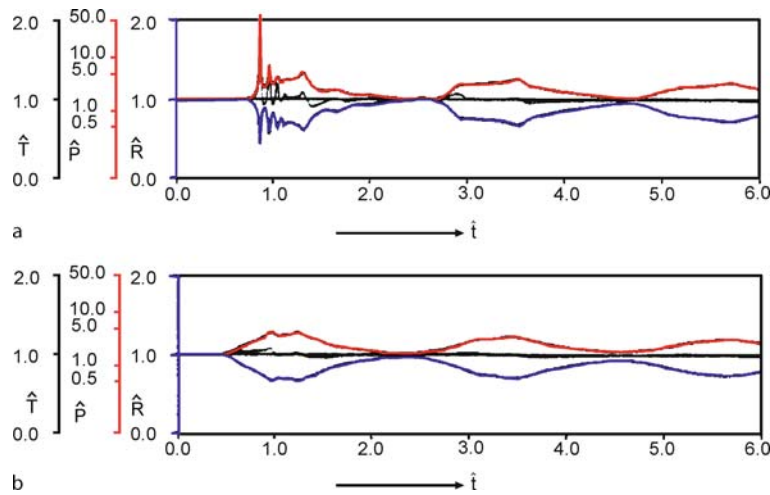
### Non-linear Oscillation of a Spherical Cloud of Bubbles

Oscillation of a group of bubbles generates particular signals as well as oscillation of a single bubble. The presence of bubbles can lower the acoustic speed of the fluid by an order of magnitude [16,42], if the liquid viscosity is small enough [30,31]. Therefore, there is a sharp impedance contrast between a region of bubbly liquid and a region of pure liquid. The boundary of a bubble cloud acts like an elastic boundary that traps acoustic energy in the bubbly region so that the bubble cloud has characteristic frequency of resonance [56,67,102]. Chouet [10] considered that it is the mechanism for the harmonic oscillations observed at volcanoes having relatively low-viscosity magma. He showed the typical values for the oscillation in the few Hz range may be generated by a columnar bubble cloud



Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 14

Growth of a vapor bubble by rectified heat transfer. Bubble radius normalized by the linear resonant radius  $R_r = 2.71$  mm versus time for saturate water at 1 atm. The sound frequency is 1 kHz and the amplitude is attached to each profile (Fig. 4 in [24])



**Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 15**

Change of bubble radius (blue lines) and pressure (red lines) in a spherical cloud of bubbles after a sudden pressure rise. The initial pressure and temperature are  $3 \times 10^4$  Pa, 293 K, radii of the bubble and the cloud are  $2.5 \times 10^{-4}$  m and  $3 \times 10^{-2}$ , the void fraction is 3%, and the pressure rise is  $2.4 \times 10^4$  Pa. Values at the center (a) and at a half the radius from the center (b) are presented in dimensionless form (a unit of the dimensionless time corresponds to  $\sim$  ms). (Modified from Fig. 9 in [67])

with void fraction of 1%,  $\sim 100$  m in length, and  $\sim 1$  m in radius in a bubble-free magma with  $\sim 5$  m in radius. The radius of each bubble is assumed as  $10^{-3}$  m, which has its own resonance frequency at 8.9 kHz [10].

Omta [67] conducted numerical calculation for oscillation of a spherical cloud of bubbles with relatively large amplitude. Figure 15 shows one of his results, in which the cloud oscillation was excited by an external pressure increase. Three cycles of the oscillation are presented. We can see high-frequency strong pulsation at the center of the bubble cloud (Fig. 15a). It is explained as follows [67]. The pressure perturbation is amplified and sharpened toward the center of the cloud because of the spherical geometry. Then the bubbles are excited at their resonance frequency.

The strong pulsation is observed only near the center in a spherical bubble cloud (Fig. 15). If a bubble cloud is hemispherical with its cross section attached on a solid wall, the pulsation generated at the center of the hemisphere may strongly hit the wall [86]. Generation of strong high-frequency pulses by a bubble cloud interacting with a pressure perturbation with lower frequencies is actually observed in experiments and now the phenomenon is going to be applied to medical treatment under controlled condition [33,60]. Similar mechanisms may work in a hydrothermal system, in which a low-frequency perturbation generates strong pressure impulses hitting the walls to be observed as seismic waves. In fact pressure oscillations in the bubble cloud (Fig. 15) and in the geyser (Fig. 11) have

quite similar features, though their time scales are different by three orders of magnitude.

### Future Directions

We have summarized theoretical bases of the bubble dynamics, mainly based on radial motion equations of a single bubble. These theories have been established and verified by experiments for simple systems. For volcanic systems, these theories have mainly been applied to the nucleation and growth of bubbles in magma. This subject takes an important role of volcanology, though it is not included in this review paper. By comparing the theory with observation of bubbles left in natural volcanic rock [58, 94] and re-producing the process by laboratory experiments [23,52,90], researchers have determined physical parameters of the volcanic processes, which are the temperature, pressure, volatile saturation, ascent rate, and so on. The bubble dynamics theories have also been applied to explain geophysical observations, as we have reviewed several possible mechanisms of bubbles that generate pressure impulses. However, determining the effects or even existence of bubbles is more difficult in these phenomena than in the bubble growth problems, because direct observation of bubbles and re-production of the process in a laboratory are more difficult. Here we discuss how we can go forward to confirm the models and apply them to determine useful physical parameters.

It would be effective to focus on relatively simple volcanic phenomena in which the bubble dynamics theory appears to work. Especially, for some volcanoes which erupt frequently, data taken by modern geophysical methods are being accumulated and the phenomenological cause-and-result relations between an eruption and pressure impulses before and during the eruption are well documented. For example, at Stromboli Volcano all the sequences of the repetitive small eruptions and a few proximal explosions have been taken by multi-parameter monitoring systems [84,85]. Geophysical data taken close to active craters at Sakurajima, Suwanosejima, and Semeru volcanoes have revealed common features in the pressure change before and during an explosion [32]. At Sakurajima Volcano, behaviors of a shallow gas pocket are discussed in the sequences of seismic and explosion events based on analyses of seismic data [91,92]. It might be possible and useful to have a common backbone model for these eruptions, based on which we can explain the particular detail of each case as a result of different parameters of the system.

The bubble dynamics theory which is used in the models needs to be updated, too. Responses of a bubbly fluid are sensitive to the system parameters, which determine possibilities, features, and time scales of the individual mechanisms, as are shown in the text. Although many models assume uniform system for simplicity, the natural system is considered to be non-uniform. In other words, regions having different physical parameters may coexist. Interaction of these subsystems may enhance the characteristic response, but may diminish with one another, or generate completely different effects. The inhomogeneities and their interactions may occur in various scales and manners. For example, in Sect. “[Bubbly Magma in an Elastic Rock as a Pressure Source](#)”, behaviors of a single uniform magma body in an elastic rock have been discussed. In the real system, the bubble size and the chemical composition are likely to be non-uniform within a small region and/or over the entire magma body. If the system is large, the hydrostatic pressure gradient is significant. Moreover, if there are multiple magma containers which have different physical parameters, each magma body will respond to the pressure perturbation differently, and pressure gradient may be generated between the two adjacent containers. Developing a model including these subscale interactions may be a subject of modern multi-scale multi-physics studies.

Laboratory experiments using analogous materials are useful in verifying and improving the models. In the procedure to construct a model system, we frequently find factors which are important in the real system but have been

neglected in the idealized mathematical model. Although there may be some processes which can be realized in the nature more easily, and there always be a scaling problem, experiments will give us more concrete idea about the mechanism of the models. Although most of the previous analogue experiments are designed to be compared with geological and petrological observations, there are some which investigate generation of pressure impulses by bubbles and are intended to explain seismic and/or acoustic observations [34,35,83]. According to the results and implications obtained by these preceding works, laboratory studies in this direction are promising.

It is also important to connect geophysical observation and geological data to understand the bubble dynamics phenomena in volcanology. Compared with other geophysical processes which occur in the earth, there is larger possibility that the source of activity appears to the surface after relatively short time. The geological samples (e.g., pyroclasts during eruptions and volcanic gasses) can inform us physical and chemical properties of the materials generating pressure impulses, and would make useful constraints on the model. On the other hand, these constraints can be tested and verified by geophysical signals of seismic, geodetic and acoustic measurements when the models are established.

By combining these theoretical, experimental, geophysical, and geological approaches, we will get better understanding on the processes which generate volcanic activities.

## Acknowledgments

The authors are grateful to Dr. B. Chouet, and Dr. H. Kawashima for useful information and advise. We also thank Dr. M. Kameda and two anonymous reviewers for their help to improve the manuscript.

## Bibliography

1. Aljishi S, Tatarkievicz J (1991) Why does heating water in a kettle produce sound? *Am J Phys* 59:628–632
2. Barclay J, Riley DS, Sparks RSJ (1995) Analytical models for bubble growth during decompression of high viscosity magmas. *Bull Volcanol* 57:422–431
3. Benoit JP, McNutt SR (1997) New constraints on source processes of volcanic tremor at Arenal volcano, Costa Rica, using broadband seismic data. *Geophys Res Lett* 24:449–452
4. Blower JD, Mader HM, Wilson SDR (2001) Coupling of viscous and diffusive controls on bubble growth during explosive volcanic eruptions. *Earth Planet Sci Lett* 193:47–56
5. Bowers TS (1995) Pressure-volume-temperature properties of H<sub>2</sub>O CO<sub>2</sub> fluids. In: Ahrens TJ (ed) *A Handbook of Physical Constants: Rock Physics and Phase Relations*. AGU Reference Shelf Series 3, AGU, pp 45–72

6. Brodsky EE, Sturtevant B, Kanamori H (1998) Earthquakes, volcanoes, and rectified diffusion. *J Geophys Res* 103:23827–23838
7. Brodsky EE, Karakostas V, Kanamori H (2000) A new observation of dynamically triggered regional seismicity: Earthquakes in Greece following the august, 1999 Izmit, Turkey earthquake. *Geophys Res Lett* 27:2741–2744
8. Brodsky EE, Roeloffs E, Woodcock D, Gall I, Manga M (2003) A mechanism for sustained groundwater pressure changes induced by distant earthquakes. *J Geophys Res* 108(B8):2390. doi: 10.1029/2002JB002321
9. Campos FB, Lage PLC (2000) Heat and mass transfer modeling during the formation and ascension of superheated bubbles. *Int J Heat Mass Transf* 43:2883–2894
10. Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling R (eds) *Monitoring and Mitigation of Volcano Hazards*. Springer, Berlin, pp 23–97
11. Chouet B, Dawson P, Ohminato T, Martini M (2003) Source mechanisms of explosions at Stromboli volcano, Italy, determined from moment-tensor inversions of very-long-period data. *J Geophys Res* 108
12. Chouet B, Dawson P, Arciniega-Ceballos A (2005) Source mechanism of vulcanian degassing at Popocatepetl volcano, Mexico, determined from waveform inversions of very long period signals. *J Geophys Res* 110:B07301
13. Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. *J Geophys Res* 111: B07310
14. Cole RH (1948) *Underwater Explosions*. Dover, New York
15. Collier L, Neuberg JW, Lensky N, Lyakhovsky V, Navon O (2006) Attenuation in gas-charged magma. *J Volcanol Geotherm Res* 153:21–36
16. Commander KW, Prosperetti A (1989) Linear pressure waves in bubbly liquid – comparison between theory and experiments. *J Acoust Soc Am* 85:732–746
17. Doinikov AA (2005) Equations of coupled radial and translational motions of a bubble in a weakly compressible liquid. *Phys Fluids* 17:128101
18. Eller A, Flynn HG (1965) Rectified diffusion during non-linear pulsations of cavitation bubbles. *J Acoust Soc Am* 37:493–503
19. Finch RD, Neppiras EA (1973) Vapor bubble dynamics. *J Acoust Soc Am* 53:1402–1410
20. Florschuetz LW, Chao BT (1965) On the mechanics of vapor bubble collapse. *Trans ASME, J Heat Transf* 87:209–220
21. Fogler HS, Goddard JD (1970) Collapse of spherical cavities in viscoelastic fluids. *Phys Fluids* 13:1135–1141
22. Garces MA, McNutt SR (1997) Theory of the airborne sound field generated in a resonant magma conduit. *J Volcanol Geotherm Res* 78:155–178
23. Gardner JE, Hilton M, Carroll MR (1999) Experimental constraints on degassing of magma: isothermal bubble growth during continuous decompression from high pressure. *Earth Planet Sci Lett* 168:201–218
24. Hao Y, Prosperetti A (1999) The dynamics of vapor bubbles in acoustic pressure fields. *Phys Fluids* 11:2008–2019
25. Hao Y, Prosperetti A (2002) Rectified heat transfer into translating and pulsating vapor bubbles. *J Acoust Soc Am* 112:1787–1796
26. Holloway JR, Blank JG (1994) Application of experimental results to C-O-H species in natural melts. In: Carroll MR, Holloway JR (eds) *Volatiles in Magma*. Rev Miner, vol 30. Mineral Soc Am, Washington, pp 187–230
27. Hsieh DY, Plesset MS (1961) Theory of rectified diffusion of mass into gas bubbles. *J Acoust Soc Am* 33:206–215
28. Ichihara M (2007) Dynamics of a spherical viscoelastic shell: Implications to a criterion for fragmentation/expansion of bubbly magma. *Earth Planet Sci Lett* 265:18–32
29. Ichihara M, Brodsky EE (2006) A limit on the effect of rectified diffusion in volcanic systems. *Geophys Res Lett* 33:L02316
30. Ichihara M, Kameda M (2004) Propagation of acoustic waves in a visco-elastic two-phase system: influences of the liquid viscosity and the internal diffusion. *J Volcanol Geotherm Res* 137:73–91
31. Ichihara M, Ohkunitani H, Ida Y, Kameda M (2004) Dynamics of bubbly oscillation and wave propagation in viscoelastic liquids. *J Volcanol Geotherm Res* 129:37–60
32. Iguchi M, Tameguri T, Yakiwara H (2006) Source mechanisms of volcanic explosion revealed by geophysical observations at Sakurajima, Suwanosejima and Semeru volcanoes. *Eos Trans AGU* 87 (Fall Meet Suppl); Abstract V31G–03
33. Ikeda T, Yoshizawa S, Tosaki M, Allen JS, Takagi S, Ohta N, Kitamura T, Matsumoto Y (2006) Cloud cavitation control for lithotripsy using high intensity focused ultrasound. *Ultrasound Med Biol* 32:1383–1397
34. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. *J Volcanol Geotherm Res* 129:61–82
35. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insights into a volcano-seismic source process in low-viscosity magmas. *J Geophys Res* 111:B05201
36. Johnson JB, Aster RC, Kyle PR (2004) Triggering of volcanic eruptions. *Geophys Res Lett* 31:L14604. doi:10.1029/2004GL020020
37. Kameda M, Matsumoto Y (1996) Shock waves in a liquid containing small gas bubbles. *Phys Fluids* 8:322–335
38. Kawashima H, Ichihara M, Kameda M (2001) Oscillation of a vapor/gas bubble with heat and mass transport. *Trans JSME, Ser B* 67:2234–2242
39. Kedar S, Sturtevant B, Kanamori H (1996) The origin of harmonic tremor at old faithful geyser. *Nature* 379:708–711
40. Kedar S, Kanamori H, Sturtevant B (1998) Bubble collapse as the source of tremor at old faithful geyser. *J Geophys Res* 103:24283–24299
41. Keller JB, Kolodner II (1956) Damping of underwater explosion bubble oscillations. *J Appl Phys* 27:1152–1161
42. Kieffer SW (1977) Sound speed in liquid-gas mixtures: water-air and water-steam. *J Geophys Res* 82:2895–2904
43. Kieffer SW (1989) Geologic nozzles. *Rev Geophysics* 27:3–38
44. Kumagai H, Chouet BA (2000) Acoustic properties of a crack containing magmatic or hydrothermal fluids. *J Geophys Res* 105:25493–25512
45. Kumagai H, Chouet BA (2001) The dependence of acoustic properties of a crack on the resonance mode and geometry. *Geophys Res Lett* 28:3325
46. Kumagai H, Chouet BA, Nakano M (2002) Temporal evolution of a hydrothermal system in Kusatsu–Shirane volcano, Japan,



- inferred from the complex frequencies of long-period events. *J Geophys Res* 107:2236
47. La Femina PC, Connor CB, Hill BE, Strauch W, Saballos JA (2004) Magma-tectonic interactions in Nicaragua: the 1999 seismic swarm and eruption of Cerro Negro volcano. *J Volcanol Geotherm Res* 137:187–199
  48. Landau LD, Lifshitz EM (1986) *Theory of Elasticity*, 3rd edn. Butterworth, Oxford
  49. Landau LD, Lifshitz EM (1987) *Fluid Mechanics*, 2nd edn. Pergamon Press, Oxford
  50. Lensky NG, Lyakhovsky V, Navon O (2001) Radial variations of melt viscosity around growing bubbles and gas overpressure in vesiculating magmas. *Earth Planet Sci Lett* 186:1–6
  51. Lensky NG, Lyakhovsky V, Navon O (2002) Expansion dynamics of volatile-supersaturated liquids and bulk viscosity of bubbly magmas. *J Fluid Mech* 460:39–56
  52. Lensky NG, Navon O, Lyakhovsky V (2004) Bubble growth during decompression of magma: experimental and theoretical investigation. *J Volcanol Geotherm Res* 129:7–22
  53. Lensky NG, Niebo RW, Holloway JR, Lyakhovsky V, Navon O (2006) Bubble nucleation as a trigger for xenolith entrapment in mantle melts. *Earth Planet Sci Lett* 245:278–288
  54. Linde AT, Sacks I (1998) Triggering of volcanic eruptions. *Nature* 395:888–890
  55. Linde AT, Sacks I, Johnston MJS, Hill DP, Bilham RG (1994) Increased pressure from rising bubbles as a mechanism for remotely triggered seismicity. *Nature* 371:408–410
  56. Lu NQ, Prosperetti A, Yoon SW (1990) Underwater noise emissions from bubble clouds. *IEEE J Ocean Eng* 15:275–281
  57. Manga M, Brodsky E (2006) Seismic triggering of eruptions in the far field: Volcanoes and geysers. *Ann Rev Earth Planet Sci* 34:263–291
  58. Mangan MT, Cashman KV (1996) The structure of basaltic scoria and reticulite and inferences for vesiculation, foam formation, and fragmentation in lava fountains. *J Volcanol Geotherm Res* 73:1–18
  59. Matsumoto Y, Takemura F (1994) Influence of internal phenomena on gas bubble motion (effects of thermal diffusion, phase change on the gas-liquid interface and mass diffusion between vapor and noncondensable gas in the collapsing phase). *JSME Int J, Ser B* 37:288–296
  60. Matsumoto Y, Allen JS, Yoshizawa S, Ikeda T, Kaneko Y (2005) Medical ultrasound with microbubbles. *Exp Therm Fluid Sci* 29:255–265
  61. Nakoryakov VE, Pokusaev BG, Shreiber IR (1993) *Wave Propagation in Gas-Liquid Media*, 2nd edn. CRC Press, Boca Raton
  62. Navon O, Lyakhovsky V (1998) Vesiculation processes in silicic magmas. In: Gilbert JS, Sparks RSJ (eds) *The Physics of Explosive Volcanic Eruption*. Geol Soc, Special Publications, 145, London, pp 27–50
  63. Nigmatulin RI, Khabeev NS, Nagiev FB (1981) Dynamics, heat and mass-transfer of vapor-gas bubbles in a liquid. *Int J Heat Mass Trans* 24:1033–1044
  64. Nishimura T (2004) Pressure recovery in magma due to bubble growth. *Geophys Res Lett* 31:L12613
  65. Nishimura T, Ichihara M, Ueki S (2006) Investigation of the Onikobe geyser, NE Japan, by observing the ground tilt and flow parameters. *Earth Planets Space* 58:e21–e24
  66. Ohl CD, Tijink A, Prosperetti A (2003) The added mass of an expanding bubble. *J Fluid Mech* 482:271–290
  67. Omta R (1987) Oscillations of a cloud of bubbles of small and not so small amplitude. *J Acoust Soc Am* 82:1018–1033
  68. Oura A, Yoshida S, Kudo K (1992) Rupture process of the Ito-Oki Japan earthquake of 1989 July 9 and interpretation as a trigger of volcanic-eruption. *Geophys J Int* 109:241–248
  69. Plesset MS (1949) The dynamics of cavitation bubbles. *J Appl Mech* 16:277–282
  70. Plesset MS, Prosperetti A (1977) Bubble dynamics and cavitation. *Ann Rev Fluid Mech* 9:145–185
  71. Poritsky H (1952) The collapse or growth of a spherical bubble or cavity in a viscous fluid. *Proc First Nat Cong Appl Mech* 813–821
  72. Prosperetti A (1982) A generalization of the rayleigh-plesset equation of bubble dynamics. *Phys Fluids* 25:409–410
  73. Prosperetti A (1984) Acoustic cavitation series: part three, bubble phenomena in sound fields: part two. *Ultrasonics* 22:115–124
  74. Prosperetti A (1984) Acoustic cavitation series: part two, bubble phenomena in sound fields: part one. *Ultrasonics* 22:69–78
  75. Prosperetti A (1991) The thermal behavior of oscillating gas bubbles. *J Fluid Mech* 222:587–616
  76. Prosperetti A (2004) Bubbles. *Phys Fluids* 16:1852–1865
  77. Prosperetti A, Hao Y (2002) Vapor bubbles in flow and acoustic fields. *Ann NY Acad Sci* 974:328–347
  78. Prosperetti A, Lezzi A (1986) Bubble dynamics in a compressible liquid, 1. 1st-order theory. *J Fluid Mech* 168:457–478
  79. Prousevitch AA, Sahagian DL, Anderson AT (1993) Dynamics of diffusive bubble growth in magmas: Isothermal case. *J Geophys Res* 98:22283–22307
  80. Rayleigh L (1917) On the pressure developed in a liquid during the collapse of a spherical cavity. *Philos Mag* 34:94–98
  81. Ripepe M, Gordeev E (1999) Gas bubble dynamics model for shallow volcanic tremor at Stromboli. *J Geophys Res* 104:10639–10654
  82. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. *Geophys Res Lett* 23:181–184
  83. Ripepe M, Ciliberto S, Della Schiava M (2001) Time constraints for modeling source dynamics of volcanic explosions at Stromboli. *J Geophys Res* 106:8713–8727
  84. Ripepe M, Marchetti E, Poggi P, Harris A, Fiaschi AJL, Olivieri G (2004) Seismic, acoustic and thermal network monitors the 2003 eruption of Stromboli volcano. *EOS, Trans AGU* 85:329
  85. Ripepe M, Marchetti E, Olivieri G, DelleDonne D, Genco R, Laccana G (2007) Monitoring the 2007 Stromboli effusive eruption by an integrated geophysical network. *The 21st Century COE Earth Sci Int Symp Abstr Z* 327
  86. Shimada M, Matsumoto Y, Kobayashi T (2006) Influence of the nuclei size distribution on the collapsing behavior of the cloud cavitation. *JSME Int J Ser B* 155:307–322
  87. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. *J Volcanol Geotherm Res* 155:307–322
  88. Sparks RSJ (1978) Dynamics of bubble formation and growth in magmas – review and analysis. *J Volcanol Geotherm Res* 3:1–37
  89. Sturtevant B, Kanamori H, Brodsky EE (1996) Seismic triggering by rectified diffusion in geothermal systems. *J Geophys Res* 101:25269–25282



90. Suzuki Y, Gardner JE, Larsen JF (2007) Experimental constraints on syneruptive magma ascent related to the phreatomagmatic phase of the 2000AD eruption of Usu volcano, Japan. *Bull Volcanol* 69:423–444
91. Tameguri T, Iguchi M, Ishihara K (2002) Mechanism of explosive eruptions from moment tensor analyses of explosion earthquakes at Sakurajima volcano. *Bull Volcanol Soc Japan* 47:197–215
92. Tameguri T, Maryanto S, Iguchi M (2007) Source mechanisms of harmonic tremors at Sakurajima volcano. *Bull Volcanol Soc Japan* 52:273–279
93. Toramaru A (1995) Numerical study of nucleation and growth of bubbles in viscous magmas. *J Geophys Res* 100:1913–1931
94. Toramaru A (2006) BND (bubble number density) decompression rate meter for explosive volcanic eruptions. *J Volcanol Geotherm Res* 154:303–316
95. Vergnolle S, Brandeis G (1994) Origin of the sound generated by Strombolian explosions. *Geophys Res Lett* 21:1959–1962
96. Vergnolle S, Brandeis G (1996) Strombolian explosions. 1. A large bubble breaking at the surface of a lava column as a source of sound. *J Geophys Res* 101:20433–20447
97. Voight B, Linde AT, Sacks IS, Mattioli GS, Sparks RSJ, Elsworth D, Hidayat D, Malin PE, Shalev E, Widiwijayanti C, Young SR, Bass V, Clarke A, Dunkley P, Johnston W, McWhorter N, Neuberg J, Williams P (2006) Unprecedented pressure increase in deep magma reservoir triggered by lava-dome collapse. *Geophys Res Lett* 33:L03,312
98. Wang T (1974) Rectified heat transfer. *J Acoust Soc Am* 56:1131–1143
99. Webb SL (1997) Silicate melts: Relaxation, rheology, and the glass transition. *Rev Geophys* 35:191–218
100. Yamada K, Emori H, Nakasawa N (2006) Bubble expansion rates in viscous compressible liquid. *Earth Planets Space* 58:865–872
101. Yang B, Prosperetti A, Takagi S (2003) The transient rise of a bubble subject to shape or volume changes. *Phys Fluids* 15:2640–2648
102. Yoon SW, Crum LA, Prosperetti A, Lu NQ (1991) An investigation of the collective oscillations of a bubble cloud. *J Acoust Soc Am* 89:700–706
103. Zuber N (1961) The dynamics of vapor bubbles in nonuniform temperature fields. *Int J Heat Mass Transf* 2:83–98

## Principal-Agent Models

INÉS MACHO-STADLER, DAVID PÉREZ-CASTRILLO  
Universitat Autònoma de Barcelona, Barcelona, Spain

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Base Game](#)

[Moral Hazard](#)

[Adverse Selection](#)

### Future Directions

### Bibliography

### Glossary

**Information economics** Information economics studies how information and its distributions among the players affect economic decisions.

**Asymmetric information** In a relationship or a transaction, there is asymmetric information when one party has more or better information than other party concerning relevant characteristics of the relationship or the transaction. There are two types of asymmetric information problems: Moral Hazard and Adverse Selection.

**Principal-agent** The principal-agent model identifies the difficulties that arise in situations where there is asymmetric information between two parties and finds the best contract in such environments. The “principal” is the name used for the contractor while the “agent” corresponds to the contractee. Both principal and agent could be individuals, institutions, organizations, or decision centers. The optimal solutions propose mechanisms that try to align the interests of the agent with those of the principal, such as piece rates or profit sharing; or that induce the agent to reveal the information, such as self-reporting contracts.

**Moral hazard (hidden action)** The term moral hazard initially referred to the possibility that the redistribution of risk (such as insurance which transfers risk from the insured to the insurer) changes people’s behavior. This term, which has been used in the insurance industry for many years, was studied first by Kenneth Arrow.

In principal-agent models, the term moral hazard is used to refer to all environments where the ignorant party lacks information about the behavior of the other party *once* the agreement has been signed, in such a way that the asymmetry arises *after* the contract is settled.

**Adverse selection (hidden information)** The term adverse selection was originally used in insurance. It describes a situation where, as a result of private information, the insured are more likely to suffer a loss than the uninsured (such as offering a life insurance contract at a given premium may imply that only the people with a risk of dying over the average take it).

In principal-agent models, we say that there is an adverse selection problem when the ignorant party lacks information while negotiating a contract, in such a way that the asymmetry is *previous* to the relationship.

## Definition of the Subject

Principal-Agent models provide the theory of *contracts under asymmetric information*. Such a theory analyzes the characteristics of optimal contracts and the variables that influence these characteristics, according to the behavior and information of the parties to the contract. This approach has a close relation to *game theory* and *mechanism design*: it analyzes the strategic behavior by agents who hold private information and proposes mechanisms that minimize the inefficiencies due to such strategic behavior. The costs incurred by the principal (the contractor) to ensure that the agents (the contractees) will act in her interest are some type of *transaction cost*. These costs include the tasks of investigating and selecting appropriate agents, gaining information to set performance standards, monitoring agents, bonding payments by the agents, and residual losses.

Principal-agent theory (and *Information Economics* in general) is possibly the area of economics that has evolved the most over the past twenty-five years. It was initially developed in parallel with the new economics of *Industrial Organization* although its applications include now almost all areas in economics, from finance and political economy to growth theory.

Some early papers centered on incomplete information in insurance contracts, and more particularly on moral hazard problems, are Spence and Zeckhauser [88] and Ross [81]. The theory soon generalized to dilemmas associated with contracts in other contexts [38,46]. It was further developed in the mid-seventies by authors such as Pauly [72,73], Mirrlees [66], Harris and Raviv [39], and Holmström [40]. Arrow [7] worked on the analysis of the optimal incentive contract when the agent's effort is not verifiable.

A particular case of adverse selection is the one where the type of the agent relates to his valuation of a good. Asymmetric information about buyers' valuation of the objects sold is the fundamental reason behind the use of auctions. Vickrey [92] provides the first formal analysis of the first and second-prize auctions. Akerlof [3] highlighted the issue of adverse selection in his analysis of the market for second-hand goods. Further analyzes include the early work of Mirrlees [65], Spence [87], Rothschild and Stiglitz [83], Mussa and Rosen [68], as well as Baron and Myerson [10], and Guesnerie and Laffont [36].

The importance of the topic has also been recognized by the Nobel Foundation. James A. Mirrlees and William Vickrey were awarded with the Nobel Prize in Economics in 1996 "for their fundamental contributions to the economic theory of incentives under asymmetric informa-

tion". Five years later, in 2001, George A. Akerlof, A. Michael Spence and Joseph E. Stiglitz also obtained the Nobel Prize in Economics "for their analyzes of markets with asymmetric information".

## Introduction

The objective of the Principal-Agent literature is to analyze situations in which a contract is signed under asymmetric information, that is, when one party knows certain relevant things of which the other party is ignorant. The simplest situation concerns a bilateral relationship: the contract between one principal and one agent. The objective of the contract is for the agent to carry out actions on behalf of the principal; and to specify the payments that the principal will pass on to the agent for such actions.

In the literature, it is always assumed that the principal is in charge of designing the contract. The agent receives an offer and decides whether or not to sign the contract. He will accept it whenever the utility obtained from it is greater than the utility that the agent would get from not signing. This utility level that represents the agent's outside opportunities is his *reservation utility*. In order to simplify the analysis, it is assumed that the agent cannot make a counter offer to the principal. This way of modeling implicitly assumes that the principal has all the bargaining power, except for the fact that the reservation utility can be high in those cases where the agent has excellent outside opportunities.

If the agent decides not to sign the contract, the relationship does not take place. If he does accept the offer, then the contract is implemented. It is crucial to notice that the contract is a reliable promise by both parties, stating the principal and agent's obligations for all (contractual) contingencies. It can only be based on *verifiable variables*, that is, those for which it is possible for a third party (a court) to verify whether the contract has been fulfilled. When some players know more than others about relevant variables, we have a situation with asymmetric information. In this case, *incentives* play an important role.

Given the description of the game played between principal and agent, we can summarize its timing in the following steps:

- (i) The principal designs the contract (or set of contracts) that she will offer to the agent, the terms of which are not subject to bargaining.
- (ii) The alternatives opened to the agent are to accept or to reject the contract. The agent accepts it if he desires so, that is, if the contract guarantees him greater expected utility than any other (outside) opportunities available to him.

- (iii) The agent carries out an action or effort on behalf of the principal.
- (iv) The outcome is observed and the payments are done.

From these elements, it can be seen that *the agent's objectives may be in conflict with those of the principal*. When the information is asymmetric, the informed party tries to take advantage, while the uninformed player tries to control this behavior via the contract. Since a Principal-Agent problem is a sequential game, the solution concept to use is *Subgame (Bayesian) Perfect Equilibrium*.

The set-up gives rise to three possible scenarios:

1. The *Symmetric Information* case, where the two players share the same information, even if they both may ignore some important elements (some elements may be uncertain).
2. The *Moral Hazard* case, where the asymmetry of information arises once the contract has been signed: the decision or the effort of agent is not verifiable and hence it cannot be included in the contract.
3. The *Adverse Selection* case, where the asymmetry of information is previous to the signature of the contract: a relevant characteristic of the agent is not verifiable and hence the principal cannot include it in the contract.

To see an example of moral hazard, consider a laboratory or research center (the principal) that contracts a researcher (the agent) to work on a certain project. It is difficult for the principal to distinguish between a researcher who is thinking about how to push the project through, and a researcher who is thinking about how to organize his evening. It is precisely this difficulty in controlling effort inputs, together with the inherent uncertainty in any research project, what generates a moral hazard problem, which is a non-standard labor market problem.

For an example of adverse selection, consider a regulator who wants to set the price of the service provided by a public monopoly equal to the average costs in the firm (to avoid subsidies). This policy (as many others) is subject to important informational requirements. It is not enough that the regulator asks the firm to reveal the required information in order to set the adequate price, since the firm would attempt to take advantage of the information. Therefore, the regulator should take this problem into account.

### The Base Game

Consider a contractual relationship between a principal and an agent, who is contracted to carry out a task. The relationship allows a certain *result* to be obtained, whose

monetary value will be referred to as  $x$ . For the sake of exposition, the set of possible results  $X$  is assumed to be finite,  $X = \{x_1, \dots, x_n\}$ . The final result depends on the *effort* that the agent devotes to the task, which will be denoted by  $e$ , and the value of a random variable for which both participants have the same prior distribution. The probability of result  $x_i$  conditional on effort  $e$  can be written as:

$$\text{Prob}[x = x_i|e] = p_i(e), \text{ for } i \in \{1, 2, \dots, n\},$$

with  $\sum_{i=1}^n p_i(e) = 1$ . Let us assume that  $p_i(e) > 0$  for all  $e, i$ , which implies that no result can be ruled out for any given effort level.

The Base Game is the reference situation, where principal and agent have the same information (even the one concerning the random component that affects the result). Since uncertainty exists, participants react to risk. Risk preferences are expressed by the shape of their *utility functions* (of the von Neumann–Morgenstern type). The principal, who owns the result and must pay the agent, has preferences represented by the utility function

$$B(x - w),$$

where  $w$  represents the payoff made to the agent.  $B(\cdot)$  is assumed to be increasing and concave:  $B' > 0$ ,  $B'' \leq 0$  (where the primes represent, respectively, the first and second derivatives). The concavity of the function  $B(\cdot)$  indicates that the principal is either risk-neutral or risk averse.

The agent receives a monetary pay-off for his participation in the relationship, and he supplies an effort which implies some cost to him. For the sake of simplicity we represent his utility function as:

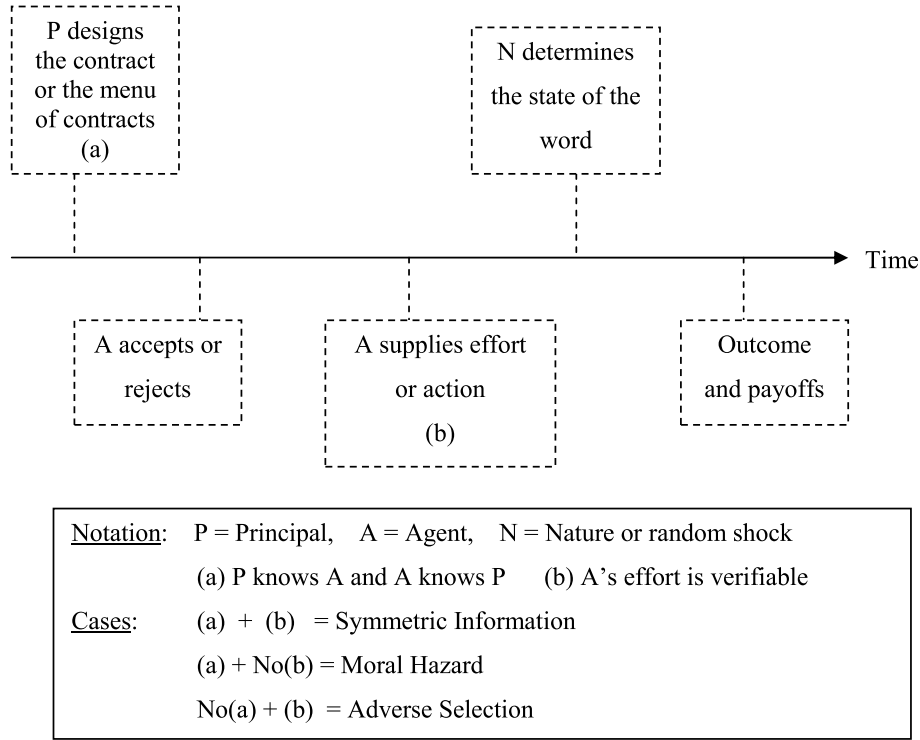
$$U(w, e) = u(w) - v(e),$$

additively separable in the components  $w$  and  $e$ . This assumption implies that the agent's risk-aversion does not vary with the effort he supplies (many results can be generalized for more general utility functions). The utility derived from the wage,  $u(w)$ , is increasing and concave:

$$u'(w) > 0, \quad u''(w) \leq 0,$$

thus the agent may be either risk-neutral,  $u''(w) = 0$ , or risk-averse  $u''(w) < 0$ . In addition, greater effort means greater disutility. We also assume that the marginal disutility of effort is not decreasing:  $v'(e) > 0$ ,  $v''(e) \geq 0$ .

A *contract* can only be based in verifiable information. In the base game, it can depend on the characteristics of the principal and the agent and includes both the effort  $e$



Principal-Agent Models, Figure 1

The figure summarizes the timing of the relationship and the three cases as a function of the information available to the participants

that the principal demands from the agent, and the wages  $\{w(x_i)\}_{i=1,\dots,n}$ .

If the agent rejects the contract, he will have to fall back on the outside opportunities that the market offers him. These other opportunities, that by comparison determine the limit for participation in the contract, are summarized in the agent's *reservation utility*, denoted by  $\underline{U}$ . So the agent will accept the contract as long as he earns an expected utility equal or higher than his reservation utility. Since, the principal's problem is to design a contract that the agent will accept, (by backward induction) the optimal contract must satisfy the *participation constraint* and it is the solution to the following maximization problem:

$$\begin{aligned} \text{Max}_{[e, \{w(x_i)\}_{i=1,\dots,n}]} \quad & \sum_{i=1}^n p_i(e) B(x_i - w(x_i)) \\ \text{s.t.} \quad & \sum_{i=1}^n p_i(e) u(w(x_i)) - v(e) \geq \underline{U}. \end{aligned}$$

The above problem corresponds to a *Pareto Optimum* in the usual sense of the term. The solution to this problem is conditional on the value of the parameter  $\underline{U}$ , so that even those cases where the agent can keep a large share of the surplus are taken into account.

The principal's program is well behaved with respect to payoffs given the assumptions on  $u(w)$ . Hence the Kuhn-Tucker conditions will be both necessary and sufficient for the global solution of the problem. However, we cannot ascertain the concavity (or quasi-concavity) of the functions with respect to effort given the assumptions on  $v(e)$ , because these functions also depend on all the  $p_i(e)$ . Hence it is more difficult to obtain global conclusions with respect to this variable.

Let us denote by  $e^\circ$  the efficient effort level. From the first-order Kuhn-Tucker conditions with respect to the wages in the different contingencies, we can analyze the associated pay-offs  $\{w^\circ(x_i)_{i=1,\dots,n}\}$ . We obtain the following condition:

$$\lambda^\circ = \frac{B'(x_i - w^\circ(x_i))}{u'(w^\circ(x_i))}, \text{ for all } i \in \{1, 2, \dots, n\},$$

where  $\lambda^\circ$  is the multiplier associated with the participation constraint. When the agent's utility is additively separable, the participation constraint binds ( $\lambda^\circ$  is positive). The previous condition equates marginal rates of substitution and indicates that the optimal distribution of risk requires that the ratio of marginal utilities of the principal and the agent to be constant irrespective of the final result.

If the principal is risk-neutral ( $B''(.) = 0$ ), then the optimal contract has to be such that  $u'(w^\circ(x_i)) = \text{constant}$  for all  $i$ . In addition, if the agent is risk-averse ( $u''(.) < 0$ ), he receives the same wage, say  $w^\circ$ , in all contingencies. This wage only depends on the effort demanded and is determined by the participation constraint. If the agent is risk-neutral ( $u''(.) = 0$ ) and the principal is risk-averse ( $B''(.) < 0$ ), then we are in the opposite situation. In this scenario, the optimal contract requires the principal's profit to be independent of the result. Consequently, the agent bears all the risk, insuring the principal against variations in the result. When both the principal and the agent are risk-averse, each of them needs to accept a part of the variability of the result. The precise amount of risk that each of them supports depends on their degrees of risk-aversion. Using the Arrow-Pratt measure of absolute risk-aversion  $r_p = -B''/B'$  and  $r_a = -u''/u'$  for the principal and the agent respectively, we can show that:

$$\frac{dw^\circ}{dx_i} = \frac{r_p}{r_p + r_a},$$

which indicates how the agent's wage changes given an increase in the result  $x_i$ . Since  $r_p/(r_p + r_a) \in (0, 1)$ , when both participants are risk-averse, the agent only receives a part of the increased result via a wage increase. The more risk-averse is the agent, that is to say, the greater is  $r_a$ , the less the result influences his wage. On the other hand, as the risk-aversion of the principal increases, greater  $r_p$ , changes in the result correspond to more important changes in the wage.

## Moral Hazard

### Basic Moral Hazard Model

Here we concentrate on *moral hazard*, which is the case in which the informational asymmetry relates to the agent's behavior during the relationship. We analyze the optimal contract when the agent's effort is not verifiable. This implies that effort cannot be contracted upon, because in case of the breach of contract, no court of law could know if the contract had really been breached or not. There are many examples of this type of situation. A traditional example is accident insurance, where it is very difficult for the insurance company to observe how careful a client has been to avoid accidents.

The principal will state a contract based on any signals that reveal information on the agent's effort. We will assume that only the result of the effort is verifiable at the end of the period and, consequently, it will be included in

the contract. However, if possible, the contract should be contingent on many other things. Any information related to the state of nature is useful, since it allows better estimations of the agent's effort thus reducing the risk inherent in the relationship. This is known as the *sufficient statistic result*, and it is perhaps the most important conclusion in the moral hazard literature [40]. The empirical content of the sufficient statistic argument is that a contract should exploit all available information in order to filter out risk optimally.

The timing of a moral hazard game is the following. In the first place, the principal decides what contract to offer the agent. Then the agent decides whether or not to accept the relationship, according to the terms of the contract. Finally, if the contract has been accepted, the agent chooses the effort level that he most desires, given the contract that he has signed. This is a free decision by the agent since effort is not a contracted variable. Hence, the principal must bear this in mind when she designs the contract that defines the relationship.

To better understand the nature of the problem faced by the principal, consider the case of a risk-neutral principal and a risk-averse agent, which implies that, under the symmetric information, the optimal contract is to completely insure the agent. However, if the principal proposes this contract when the agent's effort is not a contracted variable, once he has signed the contract the agent will exert the effort level that is most beneficial for him. Since the agent's wage does not depend on his effort, he will use the lowest possible effort.

The idea underlying an incentive contract is that the principal can make the agent interested in the consequences of his behavior by making his pay-off dependent on the result obtained. Note that this has to be done at the cost of distorting the optimal risk sharing among both participants. The trade-off between *efficiency*, in the sense of the optimal distribution of risk, and *incentives* determines the optimal contract.

Formally, since the game has to be solved by backwards induction, the *optimal contract* under moral hazard is the solution to the maximization problem:

$$\begin{aligned} & \text{Max}_{[e, \{w(x_i)\}_{i=1, \dots, n}]} \sum_{i=1}^n p_i(e) B(x_i - w(x_i)) \\ & \text{s.t. } \sum_{i=1}^n p_i(e) u(w(x_i)) - v(e) \geq \underline{U} \\ & e \in \text{Arg Max}_{\hat{e}} \left\{ \sum_{i=1}^n p_i(\hat{e}) u(w(x_i)) - v(\hat{e}) \right\} \end{aligned}$$



The second restriction is the *incentive compatibility constraint* and the first restriction is the participation constraint. The incentive compatibility constraint, and not the principal as under symmetric information, determines the effort of the agent.

The first difficulty in solving this program is related to the fact that the incentive compatibility constraint is a maximization problem. The second difficulty is that the expected utility may fail to be concave in effort. Hence, to use the first order condition of the incentive compatibility constraint may be incorrect. In spite of this, there are several ways to proceed when facing to this problem. (a) Grossman and Hart [33] propose to solve it in steps, identifying first the optimal payment mechanism for any effort and then, if possible, the optimal effort. This can be done since the problem is concave in payoffs. (b) The other possibility is to consider situations where the agent's maximization problem is well defined. One possible scenario is when the set of possible efforts is finite, in which case the incentive compatibility constraint takes the form of a finite set of inequalities. Another scenario is to write the incentive compatibility as the first-order condition of the maximization problem, and introduce assumptions that allow doing it. The last solution is known as the *first-order approach*.

Let us assume that the first-order approach is adequate, and substitute the incentive compatibility constraint in the previous program by

$$\sum_{i=1}^n p'_i(\hat{e})u(w(x_i)) - v'(\hat{e}) = 0.$$

Solving the principals program with respect to the payoff scheme, and denoting by  $\lambda$  (resp.,  $\mu$ ) the Lagrangean multiplier of the participation constraint (resp., the incentive compatibility constraint), we obtain that for all  $i$ :

$$\frac{1}{u'(w(x_i))} = \lambda + \mu \frac{p'_i(e)}{p_i(e)}.$$

This condition shows that the wage should not depend at all on the value that the principal places on the result. It depends on the results as a measure of how informative they are as to effort, in order to serve as an incentive for the agent. The wage will be increasing in the result as long as the result is increasing in effort. Hence, it is optimal that the wage will be increasing in the result only in particular cases. The necessary condition for a wage to be increasing with results is  $p'_i(e)/p_i(e)$  to be decreasing in  $i$ . In statistics, this is called the *monotonous likelihood quotient property*. It is a strong condition; for example, first-order stochastic

dominance does not guarantee the monotonous likelihood property.

### Extensions of Moral Hazard Models

The basic moral hazard setup, with a principal hiring and an agent performing effort, has been extended in several directions to take into account more complex relationships.

**Repeated Moral Hazard** Certain relationships in which a moral hazard problem occurs do not take place only once, but they are repeated over time (for example, work relationships, insurance, etc.). The duration aspect (the repetition) of the relationship gives rise to new elements that are absent in static models.

Radner [76] and Rubinstein and Yaari [84] consider infinitely repeated relationships and show that frequent repetition of the relationship allows us to converge towards the efficient solution. Incentives are not determined by the payoff scheme contingent on the result of each period, but rather on average effort, and the information available is very precise when the number of periods is large. A sufficiently threatening punishment, applied when the principal believes that the agent on average does not fulfill his task, may be sufficient to dissuade him from shirking.

When the relationship is repeated a finite number of times, the analysis of the optimal contract concentrates on different issues relating *long-term* agreements and *short-term* contracts. Note that in a repeated set up, the agent's wage and the agent's consumption in a period need not be equal. Lambert [54], Rogerson [80], and Chiappori and Macho-Stadler [15] show that long-term contracts have memory (i. e., the pay-offs in any single period will depend on the results of all previous periods) since they internalize agent's consumption over time, which depends on the sequence of payments received (as a function of the past contingencies). Malcomson and Spinnnewyn [58], Fudenberg, Holmström, and Milgrom [30], and Rey and Salanié [77] study when the optimal long-term contract can be implemented through the sequence of optimal short-term contracts. Chiappori, Macho-Stadler, Rey, and Salanié [16] show that, in order for the sequence of optimal short-term contracts to admit the same solution as the long-term contract, two conditions must be met. First, the optimal sequence of single-period contracts should have memory. That is why, when the reservation utility is invariant (is not history dependent), the optimal sequence of short-term contracts will not replicate the long-term optimum unless there exist means of smoothing consumption, that is, the

agent has access to credit markets. Second, the long-term contract must be renegotiation-proof. A contract is said to be renegotiation-proof if at the beginning of any intermediate period, no new contract or renegotiation that would be preferred by all participants is possible. When the long term contract is not renegotiation-proof (i. e., if it is not possible for participants to change the clauses of the contract at a certain moment of time even if they agree), it cannot coincide with the sequence of short term contracts.

**One Principal and Several Agents** When a principal contracts with more than one agent, the stage where agents exert their effort, which is translated into the incentive compatibility, depends on the *game among the agents*. If the agents behave as a coordinated and cooperating group, then the problem is similar to the previous one where the principal hires a team. A more interesting case appears when agents play a non-cooperative game and their strategies form a Nash equilibrium.

Holmström [40] and Mookherjee [67], in models where there is *personalized information* about the output of each agent, show that the principal is interested in paying each agent according to his own production and that of the other agents if these other results can inform on the actions of the agent at hand. Only if the results of the other agents do not add information or, in other words, if an agent's result is a *sufficient statistic* for his effort, then he will be paid according to his own result.

When the only verifiable outcome is the final result of teamwork (*joint production models*), the optimal contract can only depend on this information and the conclusions are similar to those obtained in models with only one agent. Alchian and Demsetz [5] and Holmström [41] show that joint production cannot lead to efficiency when all the income is distributed amongst the agents, i. e., if the budget constraint always binds. Another player should be contracted to control the productive agents and act as the residual claimant of the relationship.

Tirole [90] and Laffont [49] have studied the effect of coalitions among the agents in an organization on their payment scheme. If collusion is bad for the organization, it adds another dimension of moral hazard (the colluding behavior). The principal may be obliged to apply rules that are collusion-proof, which implies more constraints and simpler contracts (more bureaucratic). When coordination can improve the input of a group of agents, the optimal contract has to find payment methods that strengthen group work (see [44,56]).

Another principal's decision when she hires several agents is the organization with which she will relate. This

includes such fundamental decisions as how many agents to contract and how should they be structured. These issues have been studied by Demski and Sappington [23], Melumad and Reichelstein [63], and Macho-Stadler and Pérez-Castrillo [57].

Holmström and Milgrom [42] analyze a situation in which the agent carries out *several tasks*, each one of which gives rise to a different result. They study the optimal contract when tasks are complementary (in the sense that exerting effort in one reduces the costs of the other) or substitutes. Their model allows to build a theory of job design and to explain the relationship among responsibility and authority.

**Several Principals and One Agent** When one agent works for (or signs his contracts with) several principals simultaneously (*common agency* situation), in general, the principals are better off if they cooperate. When the principals are not able to achieve the coordination and commitment necessary to act as a single individual and they do not value the results in the same way, they each demand different efforts or actions from the agent. Bernheim and Whinston [12] show that the effort that principals obtain when they do not cooperate is less than the effort that would maximize their collective profits. However, the final contract that is offered to the agent minimizes the cost of getting the agent to choose the contractual effort.

## Adverse Selection

### Basic Adverse Selection Model

Adverse selection is the term used to refer to problems of asymmetric information that appear before the contract is signed. The classic example of Akerlof [3] illustrates very well the issue: the buyer of a used car has much less information about the state of the vehicle than the seller. Similarly, the buyer of a product knows how much he appreciates the quality, while the seller only has statistical information about a typical buyer's taste [68]; or the regulated firm has more accurate information about the marginal cost of production than the regulator.

A convenient way to model adverse selection problems is to consider that the agent can be of different *types*, and that the agent knows his type before any contract is signed while the principal does not know it. In the previous examples, the agent's type is the particular quality of the used car, the level of appreciation of quality, or the firm's marginal cost. How can the principal deal with this informational problem? Instead of offering just one contract for every (or several) types of agents, she can pro-

pose several contracts so that each type of agent chooses the one that is best for him. A useful result in this literature is the *revelation principle* [31,32,69] that states that any mechanism that the principal can design is equivalent to a *direct revelation mechanism* by which the agent is asked to reveal his type and a contract is offered according to his declaration. That is, a direct revelation mechanism offers a *menu of contracts* to the agent (one contract for each possible type), and the agent can choose any of the proposed contracts. Clearly, the mechanism must give the agent the right incentives to choose the appropriate contract, that is, it must be a *self-selection mechanism*. Menus of contracts are not unusual. For instance, insurance companies offer several possible insurance contracts between which clients may freely choose their most preferred. For example, car insurance contracts can be with or without deductible clauses. The second goes to more risk averse or more frequent drivers while deductibles attract less risk averse or less frequent drivers.

Therefore, the timing of an adverse selection game is the following. In the first place, the agent's characteristics (his "type") are realized, and only the agent learns them. Then, the principal decides the menu of contracts to offer to the agent. Having received the proposal, the agent decides which one of the contracts (if any) to accept. Finally, if one contract has been accepted, the agent chooses the predetermined effort and receives the corresponding payment.

A simple model of adverse selection is the following. A risk-neutral principal contracts an agent (who could be risk-neutral or risk-averse) to carry out some verifiable effort on her behalf. Effort  $e$  provides an expected payment to the principal of  $\Pi(e)$ , with  $\Pi'(e) > 0$  and  $\Pi''(e) < 0$ . The agent could be either of *two types* that differ with respect to the disutility of effort, which is  $v(e)$  for type G(good), and  $kv(e)$ , with  $k > 1$  for type B(bad). Hence, the agent's utility function is either  $U^G(w, e) = u(w) - v(e)$  or  $U^B(w, e) = u(w) - kv(e)$ . The principal considers that the probability for an agent to be type-G is  $q$ , where  $0 < q < 1$ .

The principal designs a menu of contracts  $\{(e^G, w^G), (e^B, w^B)\}$ , where  $(e^G, w^G)$  is directed towards the most efficient type of agent, while  $(e^B, w^B)$  is intended for the least efficient type. For the menu of contracts to be a sensible proposal, the agent must be better off by truthfully revealing his type than by deceiving the principal. The principal's problem, is therefore to maximize her expected profits subject to the restrictions that (a) after considering the contracts offered, the agent decides to sign with the principal (*participation constraints*), and (b) each agent chooses the contract designed for his particular type (*in-*

*centive compatibility constraints*):

$$\begin{aligned} & \text{Max}_{[(e^G, w^G), (e^B, w^B)]} q[\Pi(e^G) - w^G] + (1 - q)[\Pi(e^B) - w^B] \\ & \text{s.t. } u(w^G) - v(e^G) \geq \underline{U} \\ & \quad u(w^B) - kv(e^B) \geq \underline{U} \\ & \quad u(w^G) - v(e^G) \geq u(w^B) - v(e^B) \\ & \quad u(w^B) - kv(e^B) \geq u(w^G) - kv(e^G) \end{aligned}$$

The main characteristics of the optimal contract menu  $\{(e^G, w^G), (e^B, w^B)\}$  are the following:

- (i) The contract offered to the good agent  $(e^G, w^G)$  is efficient ('*non distortion at the top*'). The optimal salary  $w^G$  however is higher than under symmetric information: this type of agent receives an *informational rent*. That is, the most efficient agent profits from his private information and in order to reveal this information he has to receive a utility greater than his reservation level.
- (ii) The participation condition binds for the agent when he has the highest costs (he just receives his reservation utility). Moreover, a distortion is introduced into the efficiency condition for this type of agent. By distorting, the principal loses efficiency with respect to type-B agents, but she pays less informational rent to the G-types.

### Principals Competing for Agents in Adverse Selection Frameworks

Starting with the pioneer work by Rothschild and Stiglitz [83] on insurance markets, there have been many studies on markets with adverse selection problems where there is competition among principals to attract agents. We move from a model where one principal maximizes her profits subject to the above constraints, to a game theory environment where each principal has to take into account the actions by others when deciding which contract to offer. In this case, the adverse selection problem may be so severe that we may find ourselves in situations in which no equilibrium exists.

To highlight the main results in this type of models, consider a simple case in which there are two possible risk-averse agent types: good (G) and bad (B) with G being more productive than B. In particular, we assume that G is more careful than B, in the sense that he commits fewer errors. When the agent exerts effort, the result could be either a success (S) or a failure (F). The probability that it is successful is  $p^G$  when the agent is type-G and  $p^B$  when he

is type B, where  $p^G > p^B$ . The principal values a successful result more than a failure. The result is observable, so that the principal can pay the agent according to the result, if she so desires.

There are several risk-neutral principals. Therefore, we look for the set of equilibrium contracts in the game played by principals competing to attract agents. Equilibrium contracts must satisfy that there does not exist a principal who can offer a different contract that would be preferred by all or some of the agents and that gives that principal greater expected profits. This is why, if information was symmetric, the equilibrium contracts would be characterized by the following properties: (i) principals' expected profits are zero; and (ii) each contract must be efficient. Hence the agent receives a fixed contract insuring him against random events. In particular, the equilibrium salary that the agent receives under symmetric information is higher when he is of type G than when he is of type B.

When the principals cannot observe the type of the agent, the previous contracts cannot be longer an equilibrium: all the agents would claim to be a good type. An equilibrium contract pair  $\{C^G, C^B\}$  must satisfy the condition that no principal can add a new contract that would give positive expected profits to the agents that prefer this new contract to  $C^G$  and  $C^B$ . If the equilibrium contracts for the two agent types turn out to be the same, that is, there is only one contract that is accepted by both agent types, then the equilibrium is said to be *pooling*. On the other hand, when there is a different equilibrium contract for each agent type, then we have a *separating* equilibrium. In fact, pooling equilibria never exist, since pooling contracts always give room for a principal to propose a profitable contract that would only be accepted by the G-types (the best agents). If an equilibrium does exist, it must be such that each type of agent is offered a different contract.

If the probability that the agent is "good" is large enough, then a separating equilibrium does not exist either. That is, an adverse selection problem in a market may provoke the absence of any equilibrium in that market. When, a separating equilibria does exist. In, the results are similar to the ones under moral hazard in spite of the differences in the type of asymmetric information and in the method of solving. That is, contingent pay-offs are offered to the best agent to allow the principal to separate them from the less efficient ones. In this equilibrium, the least efficient agents obtain the same expected utility (and even sign the same contract) as under symmetric information, while the best agents lose expected utility due to the asymmetric information.

## Extensions of Adverse Selection Models

**Repeated Adverse Selection** In this extension, we consider whether the repetition of the relationship during several periods helps the principal and how it influences the form of the optimal contract. Note first that if the agent's private information is different in each period and the information is not correlated among periods, then any current information revealed does not affect the future and hence the repeated problem is equivalent to simple repetition of the initial relationship. The optimal intertemporal contract will be the sequence of optimal single-period contracts.

Consider the opposite situation where the agent's type is constant over time. If the agent decides to reveal his type truthfully in the first period, then the principal is able to design efficient contracts that extract all surpluses from the agent. Hence, the agent will have very strong incentives to misrepresent his information in the early periods of the relationships. In fact, Baron and Besanko [8] show that if the principal can commit herself with a contract that covers all the periods, then the optimal contract is the repetition of the optimal static contract. This implies that the contract is not *sequentially rational*, and it is also non *robust* to renegotiation: once the first period is finished, the principal "knows" the agent's type and a better contract for both parties is possible.

It is often the case that the principal cannot commit not to renegotiate a long-term contract. Laffont and Tirole [52] show that, in this case, it may be impossible to propose perfect revelation (separating) contracts in the first periods. This is known as the *ratchet effect*. Also, Freixas, Guesnerie, and Tirole [29], and Laffont and Tirole [51] have proven that, even when separating contracts exist, they may be so costly that they are often non optimal and we should expect that information be revealed progressively over time. Baron and Besanko [9] and Laffont and Tirole [53] also introduce frameworks in which it is possible to propose perfect revelation contracts but they are not optimal.

**Relationships with Several Agents: Auctions** One particularly interesting case of relationship among one principal and several agents is that of a seller who intends to sell one or several items to several interested buyers, where buyers have private information about their valuation for the item(s). A very popular selling mechanism in such a case is an *auction*. As Klemperer [48] would put it, auction theory is one of economics' success stories in both practical and theoretical terms. Art galleries generally use *English auctions*: the agents bid "upwards";

while fish markets are generally examples of *Dutch auctions*: the seller reduces the price of the good until someone stops the auction by buying. Public contracts are generally awarded through (*first price* or *second price*) *sealed-bid auctions* where buyers introduce their bid in a closed envelope, the good is sold to the highest bidder and the prize is either the winner's own bid or the second highest bid.

Vickrey [92,93] was the first to establish the key result in auction theory, the *Revenue Equivalence Theorem* which, subject to some reasonable conditions, says that the seller can expect equal profits on average from all the above (and many other) types of auctions, and that buyers are also indifferent among them all. Auctions are efficient, since the buyer who ends up with the object is the one with the highest valuation. Hence, the existence of private information does not generate any distortions with respect to who ends up getting the good, but the revenue of the seller is less than under symmetric information.

Myerson [70] solves the general mechanism design problem of a seller who wants to maximize her expected revenue, when the bidders have independent types and all agents are risk-neutral. In general, the optimal auction is more complex than the traditional English (second price) or Dutch (first price) auction. His work has been extended by many other authors. When the buyers' types are affiliated (i. e., they are not negatively correlated in any subset of their domain), Milgrom and Weber [64] show that the revenue equivalence theorem breaks down. In fact, in this situation, McAfee, McMillan, and Reny [61] show that the seller may extract the entire surplus from the bidders as if there was no asymmetric information. Starting with Maskin and Riley [60], several authors have also analyzed auctions of multiple units.

Finally, Clarke [19] and Groves [34] initiated another group of models in which the principal contracts with several agents simultaneously, but does not attempt to maximize her own profits. This is the case of the provision of a public good through a mechanism provided by a benevolent regulator.

**Relationships with Several Agents: Other Models and Organizational Design** Adverse selection models have attempted to analyze the optimal task assignment, the advantages of delegation, or the optimal structure of contractual relationships, when the principal contracts with several agents. Riordan and Sappington [79] analyze a situation where two tasks have to be fulfilled and show that if the person in charge of each task has private information about the costs associated with the task, then the assignment of tasks within the organization is an important

decision. For example, when the costs are positively correlated, then the principal will prefer to take charge of one of the phases herself while she will prefer to delegate the task when the costs are negatively correlated.

In a very general framework, Myerson [71] shows a powerful result: in adverse selection situations, centralization cannot be worse than decentralization, since it is always possible to replicate a decentralized contract with a centralized one. This result is really a generalization of the revelation principle. Baron and Besanko [11] and Melumad, Mookherjee, and Reichelstein [62] show that, if the principal can offer complex contracts in a decentralized organization, then a decentralized structure can replicate a centralized organization. When there are problems of communication between principal and agents, the equivalence result does not hold: Melumad and Reichelstein [63] show that *delegation* of authority can be preferable if communication between the principal and the agents is difficult. Still concerning the optimal design of the organization, Dana [22] analyzes the optimal hierarchical structure in industries with several productive phases, when firms have private information related to their costs. They show that structures that concentrate all tasks to a single agent are superior since, the incentives to dishonestly reveal the costs of each of the phases are weaker. Da-Rocha-Alvarez and De-Frutos [20] argue that the absolute advantage of the centralized hierarchy is not maintained if the differences in costs between the different phases are sufficiently important.

**Several Principals** Stole [89] and Martimort [59] point out the difficulty of extending the revelation principle to situations where an agent with private information is contracted by several principals who act separately. Given that not only one contract (or menu of contracts) is offered to the agent, but several contracts coming from different principals, it is not longer necessarily true that the best a principal can do is to offer a "truth-telling mechanism".

Consider a situation with two principals that are hiring a single agent. If we accept that agent's messages are restricted to the set of possible types that the agent may have, we can obtain some conclusions. If the activities or efforts that the agent carries out for the two principals are substitutes (for instance, a firm producing for two different customers), then the usual result on the distortion of the decision holds: the most efficient type of agent supplies the efficient level of effort while the effort demanded from the least efficient type is distorted. However, due to the lack of cooperation between principals, the distortion induced in the effort demanded from the less efficient type of agent is lower than the one maximizing the principals'



aggregate profits. On the other hand, if the activities that the agent carries out for the principals are complementary (for example, the firm produces a final good that requires two complementary intermediate goods in the production process), then the comparison of the results under cooperation and under no cooperation between the principals reveals that: if a principal reduces the effort demanded from the agent, in the second case, this would imply that it is also profitable for the other principal to do the same. Therefore, the distortion in decisions is greater to that produced in the case in which principals cooperate.

**Models of Moral Hazard and Adverse Selection** The analysis of principal-agent models where there are simultaneously elements of moral hazard and adverse selection is a complex extension of classic agency theory. Conclusions can be obtained only in particular scenarios. One common class of models considers situations where the principal cannot distinguish the part corresponding to effort from the part corresponding to the agent's efficiency characteristic because both variables determine the production level. Picard [74] and Guesnerie, Picard, and Rey [37] propose a model with risk-neutral participants and show that, if the effort demanded from the different agents is not decreasing in the characteristic (if a higher value of this parameter implies greater efficiency), then the optimal contract is a menu of distortionary deductibles designed to separate the agents. The menu of contracts includes one where the principal sells the firm to the agent (aiming at the most efficient type), and another contract where she sells only a part of the production at a lower prize (aiming at the least efficient type). However, there are also cases where fines are needed to induce the agents to honestly reveal their characteristic.

In fact, the main message of the previous literature is that the optimal solution for problems that mix adverse selection and moral hazard does not imply efficiency losses with respect to the pure adverse selection solution when the agent's effort is observable. However, in other frameworks (see [50]), a true problem of asymmetric information appears only when both problems are mixed when, and efficiency losses are evident. Therefore, the same solution as when only the agent's characteristic is private information cannot be achieved.

## Future Directions

### Empirical Studies of Principal-Agent Models

The growing interest on empirical issues related to asymmetric information started in the mid nineties (see the survey by Chiappori and Salanie [18]). A very large part of the

literature is devoted to test the predictions of the canonical models of moral hazard and adverse selection, where there is only one dimension in which information is asymmetric. A great deal of effort is devoted to try to ascertain whether it is moral hazard, or adverse selection, or both prevalent in the market. This is a difficult task because both adverse selection and moral hazard generate the same predictions in a cross section. For instance, a positive correlation between insurance coverage and probability of accident can be due to either the intrinsically riskier drivers selecting into contracts with better coverage (as the [83] model of adverse selection will predict) or to drivers with better coverage exerting less effort to drive carefully (as the canonical moral hazard model will predict). Chiappori, Jullien, Salanié and Salanié [17] have shown that the positive correlation between coverage and risk holds more generally than in the canonical models as long as the competitive assumption is maintained.

Future empirical approaches are likely to incorporate market power (as in [17]), multiple dimensions of asymmetric information (as in [28]), as well as different measures of asymmetric information (as in [91]). These advances will be partly possible thanks to richer surveys which collect subjective information regarding agents' attributes usually unobserved by principals or agent's subjective probability distributions. The wider availability of panel data will mean that it will become easier to disentangle moral hazard from adverse selection (as in [1]). Much is to be learnt by using field experiments that allow randomly varying contract characteristics offered to individuals and hence disentangling moral hazard from adverse selection (as in [47]).

### Contracts and Social Preferences

Although principal-agent theory has proved fundamental in expanding our understanding of contract situations, real-life contracts frequently do not exactly match its predictions. Many contracts are linear and simpler, incentives are often stronger and wage gaps more compressed than expected. One possible explanation is that the theory has mainly focused on economic agents exclusively motivated by their own monetary incentives. However, this assumption leaves aside issues such as social ties, team spirit or work morale, which the human resources literature highlights. A recent strand of economic literature, known as "behavioral contract theory", has tried to incorporate social aspects into the economic modeling of contracts.

Such theory has been motivated by two types of empirical support. On the one hand, extensive interview studies with firm managers and employees [13] has shown not

only that agents care about social comparisons such as internal pay equity or effort diversity, but that their incentives to work hard are affected by them and that principals are aware of it and design their contracts accordingly. On the other hand, one of the most influential contributions of the experimental literature has been to show that, assuming that economic agents are not completely selfish (but exhibit some form of social preferences), helps organizing many laboratory data. Experiments replicating labor markets (starting with Fehr's [26]) confirm Akerlof's [4] insight that contracts may be understood as a form of gift exchange in which principals may offer a "generous" wage and agents may respond with more than the minimum effort required.

Incorporating social and psychological aspects in a systematic manner into agents' motivations has given rise to several forms of utility functions reflecting inequality aversion [14,24,27], fairness [75] and reciprocity [25]. More recently, such utility functions have been included into standard contract theory models and have helped in shortening the gap between theory predictions and real-life contracts. In particular, issues such as employees' feelings of envy or guilt towards their bosses [45], utility comparisons among employees [35,78] or peer-pressure motivating effort decisions [43,55] have proved important in widening the scope of issues principal-agent theory can help to understand.

### Principal-Agent Markets

The literature has been treating each principal-agent relation as an isolated entity. Thus, it normally takes a given relationship between a principal and an agent (or among several principals and/or several agents), and analyzes the optimal contract. In particular, the principal assumes all the bargaining power as she has the right to offer the contract she likes the most, and agent's payoff is determined by his exogenously given reservation utility. However, in markets there is typically not a single partnership but there are several. It is then interesting to consider the simultaneous determination of both the identity of the pairs that meet (i. e., the *matching* between principals and agents) and the contracts these partnerships sign. The payoffs to each principal and agent will then depend on the other principal-agent relationships being formed in the market. This analysis requires a general equilibrium-like model.

Game theory provides a very useful tool to deal with the study of markets where heterogeneous players from one side can form partnerships with heterogeneous players from the other side: the two-sided matching models. Examples of classic situations studied in two-sided matching

models (see [82,86]) are the marriage market, the college admissions model, or the assignment market (where buyers and sellers transact). Several papers extend this game theory models to situations where each partnership involves contracts and show that the simultaneous consideration of matching and contracts has important implications. Dam and Pérez-Castrillo [21] show that, in an economy where landowners contract with tenants, a government willing to improve the situation of the tenants can be interested in creating wealth asymmetries among them. Otherwise, the landowners would appropriate all the incremental money that the government is willing to provide to the agents. Serfes [85] shows that higher-risk projects do not necessarily lead to lower incentives, which is the prediction in the standard principal-agent theory, and Alonso-Paulí and Pérez-Castrillo [6] apply the theory to markets where contracts (between shareholders and managers) can include Codes of Best Practice. On the empirical side, Akerberg and Botticini [2] find strong evidence for endogenous matching between landlords and tenants and that risk sharing is an important determinant of contract choice.

Future research will extend the general equilibrium analysis of principal-agent contracts to other markets. In addition, the literature has only studied one-to-one matching models. This should be extended to situations where each principal can hire several agents, or where each agent deals with several principals. The interplay between (external) market competition and (internal) collaboration between agents or principals can provide useful insights about the characteristics of optimal contracts in complex environments.

## Bibliography

### Primary Literature

1. Abbring J, Chiappori PA, Heckman JJ, Pinquet J (2003) [Adverse Selection and Moral Hazard in Insurance: Can Dynamic Data Help to Distinguish?](#) *J European Econ Ass* 1:512–521
2. Akerberg DA, Botticini M (2002) Endogenous Matching and the Empirical Determinants of Contract Form. *J Political Econ* 110:564–592
3. Akerlof G (1970) The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism. *Quarterly J Econ* 89:488–500
4. Akerlof G (1982) Labor Contracts as a Partial Gift exchange. *Quarterly J Econ* 97:543–569
5. Alchian A, Demsetz H (1972) Production, Information Costs, and Economic Organization. *Am Econ Rev* 62:777–795
6. Alonso-Paulí E, Pérez-Castrillo D (2007) Codes of Best Practice in Competitive Markets. Mimeo
7. Arrow K (1985) The Economics of Agency. In: J Pratt, R Zeckhauser (eds) *Principals and agents: The Structure of business*. Harvard University Press, Boston

8. Baron D, Besanko D (1984) Regulation and Information in a Continuing Relationship. *Information Econ Policy* 1:267–302
9. Baron D, Besanko D (1987) Commitment and Fairness in a Dynamic Regulatory Relationship. *Rev Econ Studies* 54:413–436
10. Baron D, Myerson R (1982) Regulating a Monopoly with Unknown Costs. *Econometrica* 50:911–930
11. Baron D, Besanko D (1992) Information, Control and Organizational Structure. *J Econ Management Strategy* 1(2):237–275
12. Bernheim BD, Whinston MD (1986) Common Agency. *Econometrica* 54:923–942
13. Bewley T (1999) *Why Rewards Don't Fall During a Recession*. Harvard University Press, Cambridge
14. Bolton G, Ockenfels A (2000) ERC: A Theory of Equity, Reciprocity and Competition. *Am Econ Rev* 90:166–193
15. Chiappori PA, Macho-Stadler I (1990) Contrats de Travail Répétés: Le Rôle de la Mémoire. *Annales Econ Stat* 17:4770
16. Chiappori PA, Macho-Stadler I, Rey P, Salanié B (1994) Repeated Moral Hazard: The Role of Memory, Commitment and Access to Credit Markets. *Eur Econ Rev* 38:1527–1553
17. Chiappori PA, Jullien B, Salanié B, Salanié F (2006) *Asymmetric Information In Insurance: General Testable Implications*. *Rand J Econ* 37:783–798
18. Chiappori PA, Salanié B (2003) Testing Contract Theory: a Survey of Some Recent Work. In: Dewatripont, Hansen, Turnovsky (eds) *Advances in Economics and Econometrics*, vol. 1. Cambridge University Press, Cambridge, pp 115–149
19. Clarke E (1971) Multipart Pricing of Public Goods. *Public Choice* 11:17–33
20. Da-Rocha-Alvarez JM, De-Frutos MA (1999) A Note on the Optimal Structure of Production. *J Econ Theory* 89:234–246
21. Dam K, Pérez-Castrillo D (2006) The Principal-Agent Matching Market. *Frontiers Econ Theory. Berkeley Electro* 2(1):1–34
22. Dana JD (1993) The Organization and Scope of Agents: Regulating Multiproduct Industries. *J Econ Theory* 59:288–310
23. Demski JS, Sappington D (1986) Line-item Reporting, Factor Acquisition and Subcontracting. *J Accounting Research* 24:250–269
24. Desiraju R, Sappington D (2007) Equity and Adverse Selection. *J Econ Management Strategy* 16:285–318
25. Dufwenberg M, Kirchsteiger G (2004) A Theory of Sequential Reciprocity. *Games Econ Behavior* 47:268–298
26. Fehr E, Kirchsteiger G, Riedl A (1993) Does Fairness Prevent Market Clearing? *Quart J Econ* 108:437–460
27. Fehr E, Schmidt K (1999) A Theory of Fairness, Competition and Cooperation. *Quart J Econ* 114:817–868
28. Finkelstein A, McGarry K (2006) *Multiple Dimensions of Private Information: Evidence from the Long-term Care Insurance Market*. *Am Econ Rev* 96:938–958
29. Freixas X, Guesnerie R, Tirole J (1985) Planning under Information and the Ratchet Effect. *Rev Econ Studies* 52:173–192
30. Fudenberg D, Holmström B, Milgrom B (1990) Short-term Contracts and Long-term Agency Relationships. *J Econ Theory* 51:1–31
31. Gibbard A (1973) Manipulation for Voting Schemes. *Econometrica* 41:587–601
32. Green JR, Laffont JJ (1977) Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods. *Econometrica* 45:427–438
33. Grossman SJ, Hart OD (1983) An Analysis of the Principal-Agent Problem. *Econometrica* 51:7–45
34. Groves T (1973) Incentives in Teams. *Econometrica* 41:617–631
35. Grund C, Sliwka D (2005) Envy and Compassion in Tournaments. *J Econ Manag Strateg* 14:187–207
36. Guesnerie R, Laffont JJ (1984) A Complete Solution to a Class of Principal-Agent Problems with Application to the Control of a Self-Managed Firm. *J Public Econ* 25:329–369
37. Guesnerie R, Picard P, Rey P (1989) Adverse Selection and Moral Hazard with Risk Neutral Agents. *Eur Econ Rev* 33:807–823
38. Harris M, Raviv A (1978) Some Results on Incentive Contracts with Applications to Education and Employment, Health Insurance and Law Enforcement. *Am Econ Rev* 68:20–30
39. Harris M, Raviv A (1979) Optimal Incentive Contracts with Imperfect Information. *J Econ Theory* 2:231–259
40. Holmström B (1979) Moral Hazard and Observability. *Bell J Econ* 10:74–91
41. Holmström B (1982) Moral Hazard in Teams. *Bell J Econ* 13:324–340
42. Holmström B, Milgrom P (1991) Multitask Principal-Agent Analysis: Incentive Contracts, Assets Ownership, and Job Design. *J Law Econ Organization* 7(Suppl):24–52
43. Huck S, Rey-Biel P (2006) *Endogenous Leadership in Teams*. *J Inst Theoretical Econ* 162:1–9
44. Itoh H (1990) Incentives to Help in Multi-Agent Situations. *Econometrica* 59:611–636
45. Itoh H (2004) Moral Hazard and Other-Regarding Preferences. *Jpn Econ Rev* 55:18–45
46. Jensen M, Meckling W (1976) The Theory of the Firm, Managerial Behavior, Agency Costs and Ownership Structure. *J Finan Econ* 3:305–360
47. Karlan D, Zinman J (2006) *Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment*. Mimeo
48. Klemperer P (2004) *Auctions: Theory and Practice*. Princeton University Press, Princeton
49. Laffont JJ (1990) Analysis of Hidden Gaming in a Three Levels Hierarchy. *J Law Econ Organ* 6:301–324
50. Laffont JJ, Tirole J (1986) Using Cost Observation to Regulate Firms. *J Polit Econ* 94:614–641
51. Laffont JJ, Tirole J (1987) Comparative Statics of the Optimal Dynamic Incentive Contract. *Eur Econ Rev* 31:901–926
52. Laffont JJ, Tirole J (1988) The Dynamics of Incentive Contracts. *Econometrica* 56:1153–1176
53. Laffont JJ, Tirole J (1990) Adverse Selection and Renegotiation in Procurement. *Rev Econ Studies* 57:597–626
54. Lambert R (1983) Long Term Contrats and Moral Hazard. *Bell J Econ* 14:441–452
55. Lazear E (1995) *Personnel Economics*. MIT Press, Cambridge
56. Macho-Stadler I, Pérez-Castrillo D (1993) Moral Hazard with Several Agents: The Gains from Cooperation. *Int J Ind Organ* 11:73–100
57. Macho-Stadler I, Pérez-Castrillo D (1998) Centralized and Decentralized Contracts in a Moral Hazard Environment. *J Ind Econ* 46:489–510
58. Malcomson JM, Spinnewyn F (1988) The Multiperiod Principal-Agent Problem. *Rev Econ Stud* 55:391–408
59. Martimort D (1996) *Exclusive Dealing*, Common Agency and Multiprincipals Incentive Theory. *Rand J Econ* 27:1–31
60. Maskin E, Riley J (1989) Optimal Multi-Unit Auctions. In: Hahn F (ed) *The Economics of Missing Markets, Information, and Games*. Oxford University Press, Oxford, pp 312–335

61. McAfee P, McMillan J, Reny P (1989) Extracting the Surplus in the Common Value Auction. *Econometrica* 5:1451–1459
62. Melumad N, Mookherjee D, Reichstein S (1995) Hierarchical Decentralization of Incentive Contracts. *Rand J Econ* 26:654–672
63. Melumad N, Reichstein S (1987) Centralization vs Delegation and the Value of Communication. *J Account Res* 25:1–18
64. Milgrom P, Weber RJ (1982) A Theory of Auctions and Competitive Bidding. *Econometrica* 50:1089–1122
65. Mirrlees J (1971) An Exploration in the Theory of Optimum Income Taxation. *Rev Econ Studies* 38:175–208
66. Mirrlees J (1975) The Theory of Moral Hazard and Unobservable Behavior, Part I. WP Nuffield College, Oxford
67. Mookherjee D (1984) Optimal Incentive Schemes with Many Agents. *Rev Econ Studies* 51:433–446
68. Mussa M, Rosen S (1978) Monopoly and Product Quality. *J Econ Theory* 18:301–317
69. Myerson R (1979) Incentive Compatibility and the Bargaining Problem. *Econometrica* 47:61–73
70. Myerson R (1981) Optimal Auction Design. *Math Op Res* 6:58–73
71. Myerson R (1982) Optimal Coordination Mechanisms in Generalized Principal-Agent Models. *J Math Econ* 10:67–81
72. Pauly MV (1968) The Economics of Moral Hazard. *Am Econ Rev* 58:531–537
73. Pauly MV (1974) Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection. *Quart J Econ* 88:44–62
74. Picard P (1987) On the Design of Incentive Schemes under Moral Hazard and Adverse Selection. *J Public Econ* 33:305–331
75. Rabin M (1993) Incorporating Fairness into Game Theory and Economics. *Am Econ Rev* 83:1281–1302
76. Radner R (1981) Monitoring Cooperative Agreements in a Repeated Principal-Agent Relationship. *Econometrica* 49:1127–1148
77. Rey P, Salanié B (1990) Long term, Short term and Renegotiation. *Econometrica* 58:597–619
78. Rey-Biel P (2008) Inequity Aversion and Team Incentives. *ELSE WP, Scandinavian J Econ* 110:297–320
79. Riordan MH, Sappington DE (1987) Information, Incentives, and the Organizational Mode. *Quart J Econ* 102:243–263
80. Rogerson W (1985) Repeated Moral Hazard. *Econometrica* 53:69–76
81. Ross SA (1973) The Economic Theory of Agency: The Principal's Problem. *Am Econ Rev* 63:134–139
82. Roth AE, Sotomayor M (1990) Two-sided matching: A study in game-theoretic modeling and analysis. Cambridge University Press, New York
83. Rothschild M, Stiglitz J (1976) Equilibrium in Competitive Insurance Markets: An Essay in the Economics of Imperfect Information. *Quart J Econ* 90:629–650
84. Rubinstein A, Yaari ME (1983) Repeated Insurance Contrats and Moral Hazard. *J Econ Theory* 30:74–97
85. Serfes K (2008) Endogenous Matching in a Market with Heterogeneous Principals and Agents. *Int J Game Theory* 36:587–619
86. Shapley LS, Shubik M (1972) The Assignment Game I: The Core. *Int J Game Theory* 1:111–130
87. Spence M (1974) Market Signaling. Harvard University Press, Cambridge
88. Spence M, Zeckhauser R (1971) Insurance Information, and Individual Action. *Am Econ Rev* 61:380–387
89. Stole L (1991) Mechanism Design under Common Agency. WP MIT, Cambridge
90. Tirole J (1986) Hierarchies and Bureaucracies: On the Role of Collusion in Organizations. *J Law Econ Organ* 2:181–214
91. Vera-Hernandez M (2003) Structural Estimation of a Principal-Agent Model: Moral Hazard in Medical Insurance. *Rand J Econ* 34:670–693
92. Vickrey W (1961) Counterspeculation, Auctions and Competitive Sealed Tenders. *J Finance* 16:8–37
93. Vickrey W (1962) Auction and Bidding Games. In: *Recent Advances in Game Theory*. The Princeton University Conference, Princeton. pp 15–27

## Books and Reviews

- Hart O, Holmström B (1987) The Theory of Contracts. In: Bewley T (ed) *Advances in Economic Theory*, Fifth World Congress. Cambridge University Press, Cambridge
- Hirshleifer J, Riley JG (1992) *The Analytics of Uncertainty and Information*. Cambridge University Press, Cambridge
- Laffont JJ, Martimort D (2002) *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, Princeton
- Laffont JJ, Tirole J (1993) *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge
- Macho-Stadler I, Pérez-Castrillo D (1997) *An Introduction to the Economics of Information: Incentives and Contracts*. Oxford University Press, Oxford
- Milgrom P, Roberts J (1992) *Economics, Organization and Management*. Prentice-Hall, Englewood Cliffs

## Probability Densities in Complex Systems, Measuring

GUNNAR PRUESSNER

Department of Mathematics, Imperial College London, London, UK

## Article Outline

[Glossary](#)  
[Definition of the Subject](#)  
[Introduction](#)  
[Simulation Techniques](#)  
[Scaling](#)  
[Histogram Data Representation](#)  
[Future Directions](#)  
[Bibliography](#)

## Glossary

**Apparent exponent** The apparent exponent is the power-law exponent of the probability density function  $P(s; s_c)$  in the scaling region. The scaling region is the



range of event sizes  $s$  which is closest to a straight line in a double-logarithmic plot of  $P(s; s_c)$ , the intermediate range  $s_c \gg s \gg s_0$ .

**Consistent estimator** An estimator is consistent if it converges to the quantity estimated (the expectation value of the population) as the sample size is increased. For example, the lack of independence of consecutive measurements generated in a numerical simulation can render an estimator inconsistent.

**Corrections to scaling** In general, pure power-law behavior is found in an observable only to leading order, for example  $\langle s^2 \rangle = aL^\alpha + bL^\beta + \dots$  with  $\alpha > \beta$ . While the sub-leading terms can have great physical relevance more emphasis is normally given to the leading order. The quality of the data analysis when determining the leading order can improve significantly by allowing for correction terms. The exponents found in these correction terms are usually expected to be universal as well.

**Correlation time** Fitting the autocorrelation function of an observable to an exponential  $\exp(-t/\tau)$  produces the correlation time  $\tau$ . Although correlations are in general more complicated than the single exponential suggests, the standard deviation of the estimator of the  $n$ th moment from  $N$  measurements is often estimated to be  $\sqrt{(2\tau + 1)/N}$  times the estimated standard deviation of the  $n$ th moment,

$$\overline{\sigma^2}(\overline{s^n}) = \frac{2\tau + 1}{N} \overline{\sigma^2}(s^n), \quad (1)$$

as if the number of independent measurements was only  $N/(2\tau + 1)$ .

**Estimator** A numerical estimator is any function that provides an estimate from the sample, that is the set of all measurements taken. A good estimator is unbiased, consistent and efficient. Very often, such an estimator coincides with the definition of the observable as taken from the exact distribution, for example  $\overline{s^2} = \frac{1}{N} \sum_i s_i^2$  for estimating the second moment from an uncorrelated sample  $s_1, s_2, \dots, s_N$ , with exact value  $\langle s^2 \rangle = \int ds s^2 P(s; s_c)$ . However, generally, a function of observables to be estimated, is not well estimated by taking the function of the estimates. For example, the square of the first moment  $\langle s \rangle^2$  is not well estimated by the numerical estimate  $\overline{s^2} = (\sum s_i/N)^2$ , as this estimator would be biased.

**Finite size scaling** Observables in complex systems that display a power-law dependence on a parameter, often diverge in the thermodynamic limit. In finite systems they remain finite and their value is expected to

diverge as a power-law of the system size. The relation between the value of the observable and the system size is known as “finite size scaling” (abbreviated FSS).

**Gap exponent** Given that a system displays scaling, the exponents  $\gamma'_n$  characterizing the dependence of the  $n$ th moment on a parameter, such as the system size in case of finite size scaling, often are linear in  $n$ . The gap exponent is the gap between consecutive moments,  $\gamma'_{n+1} - \gamma'_n$ .

**Importance sampling** Importance sampling is a numerical technique to bias the frequency which with configurations are generated, so that states of greater importance, e.g. large observables, are generated more often than others, less important ones. Using a Markov chain to generate the states of an Ising model as opposed to generating them at random and applying a Boltzmann–Gibbs weight can be regarded as a form of importance sampling.

**Lower cutoff** The lower cutoff in a probability density function of event sizes in complex systems is a value of the event size above which the distribution displays universal behavior. Below this value the system is governed by microscopic details. For many systems, the probability density functions for different system sizes coincide for event sizes below the lower cutoff.

**Markov chain Monte Carlo** A Monte Carlo technique whereby configurations are generated by transforming the state of the system according to a transition probability. The stationary probability distribution of the different states corresponds to the target probability distribution, i.e. the distribution to be modeled. In complex systems, Markov Chain Monte Carlo (abbreviated MCMC) is the natural method to study a model: Configurations of the system are generated with the frequency they occur in the exact distribution.

**Moment analysis** In general it is a difficult to identify and quantify scaling behavior in probability density functions. The most general analysis is a data collapse, the quality of which is not easily determined. The most widely used method to determine scaling exponents and moment ratios of the universal scaling function therefore is a moment analysis. Based on the scaling assumption of the probability density function, moments scale as a power of the upper cutoff with amplitudes certain ratios of which are universal.

**Monte Carlo** Technique to calculate numerical estimates for expectation values in stochastic models by generating configurations at random. More generally, Monte Carlo (abbreviated MC) is a stochastic technique to numerically integrate a high-dimensional in-



tegral, here corresponding to the calculation of an expectation value by integrating over all degrees of freedom, constituting the phase space of the system.

**Parameter space and phase space** The parameter space of a complex system is the space spanned by all parameters of the model, such as system size and couplings of the interacting agents. A numerical study usually aims to probe the model throughout a large part of the parameter space. The number of parameters therefore needs to be as small as possible. Often only a single parameter exists. Leaving the parameters fixed, an individual numerical simulation samples the phase space available to the system. The phase space is the set of all possible configurations or states of the model. This space is very high dimensional and it is virtually impossible to sample this space homogeneously. A numerical simulation relies on the assumption that the sample taken nevertheless is sufficiently representative to allow for reliable estimates.

#### Stationary distribution and transient

Most complex systems studied possess a limiting distribution, i. e. the probability density distribution in phase space converges. This is the stationary distribution. A free random walker, for example, does not possess a stationary distribution, while a random walker in a harmonic potential does. Due to correlations, the probability distribution of states after any finite time generally depends on the initial condition, which is therefore often chosen to be random. The measurements discarded due to these correlation are called the transient.

**Unbiased estimator** An estimator is unbiased if the population average of the estimator is independent of the sample size. For example,  $\bar{s} = \sum_i^N s_i / N$  from an uncorrelated sample  $s_1, \dots, s_N$  is an unbiased estimator of the expected first moment. Estimating the variance of  $s$  as  $\bar{\sigma}^2(s) = \bar{s}^2 - \bar{s}^2$  however is biased, because the population mean of  $\bar{s}^2 - \bar{s}^2$  is  $((N-1)/N)(\langle s^2 \rangle - \langle s \rangle^2)$ , which depends on the sample size  $N$ .

**Upper cutoff** The upper cutoff is the characteristic scale of the universal part of the event size distribution. It is a measure of the event size at which the scaling function of the event size distribution breaks up. Moments  $\langle s^n \rangle$  with sufficiently large  $n$  are to leading order a power of the upper cutoff. The upper cutoff itself is expected to be a power law of the system parameter, i. e. the system size in case of finite size scaling. The exponent controlling the relation between upper cutoff and system size is the gap exponent.

#### Definition of the Subject

Observables in complex systems usually obtain a broad range of values. They are random variables of a stochastic process and the system explores a wide phase space. The observables are therefore characterized by a probability density function (PDF), representing the probability (density) to find the complex system in a state with a particular value of the observable. As in other areas of statistical mechanics, complex systems are often studied in computer simulations, most prominently **Monte Carlo** [1,2,3] and the probability density function is recorded in form of a histogram, which frequently has power-law asymptotes. Historically, the probability density function itself plays a dominant rôle in the characterization of complex system, while more recently derived quantities, in particular moments, are used more frequently.

#### Introduction

There are three main stages in the analysis of the PDF of a complex system: The first step is to generate in a computer simulation estimates of the PDF itself or derived quantities, such as moments. This is a general, technical problem considered in computational physics. In a second step, the expected behavior of the observables is to be derived, a process that often feeds to the stage of data generation, as it might suggest new observables to be estimated. Thirdly, the estimates are to be analyzed and compared to the expected behavior.

The following material highlights main aspects of these steps. In practice, the first and the second step go hand in hand and might even change places, but as the simulation techniques are so clearly distinct from the theoretical expectations and the data analysis, which, in turn, are closely related, simulation techniques are presented first.

The following subsection describes a key-example in complexity which has been analyzed in great detail using the techniques described in this article. The subsequent sections present some principal ideas and techniques from the three areas introduced above.

#### Bak–Tang–Wiesenfeld Model

The Bak–Tang–Wiesenfeld (BTW) Model [4,5] is the first model of Self Organized Criticality studied because of the peculiar features of the probability density function of its key observable. The BTW model is a cellular automaton, sometimes also called a “stochastic cellular automaton” when it is updated in a random fashion. The BTW model evolves as follows: In two dimension, height variables  $z_i$ , which can take integral values set to 0 initially, are assigned

to every site  $i$  of a square lattice of size  $L \times L$ . Whenever a height variable exceeds a certain critical value, its value is decreased by 4 and the variables of the surrounding four nearest neighboring sites are incremented by 1. These rules are modified for sites on the open boundary, where particles are lost. To drive the model, the height variable at either a single, specific site or at a randomly chosen site is increased by 1 and the relaxation rules described above are applied until none of the sites exceeds the critical value, which gives rise to an “avalanche” of size  $s$ , measuring the number of times the updating rule has been applied.

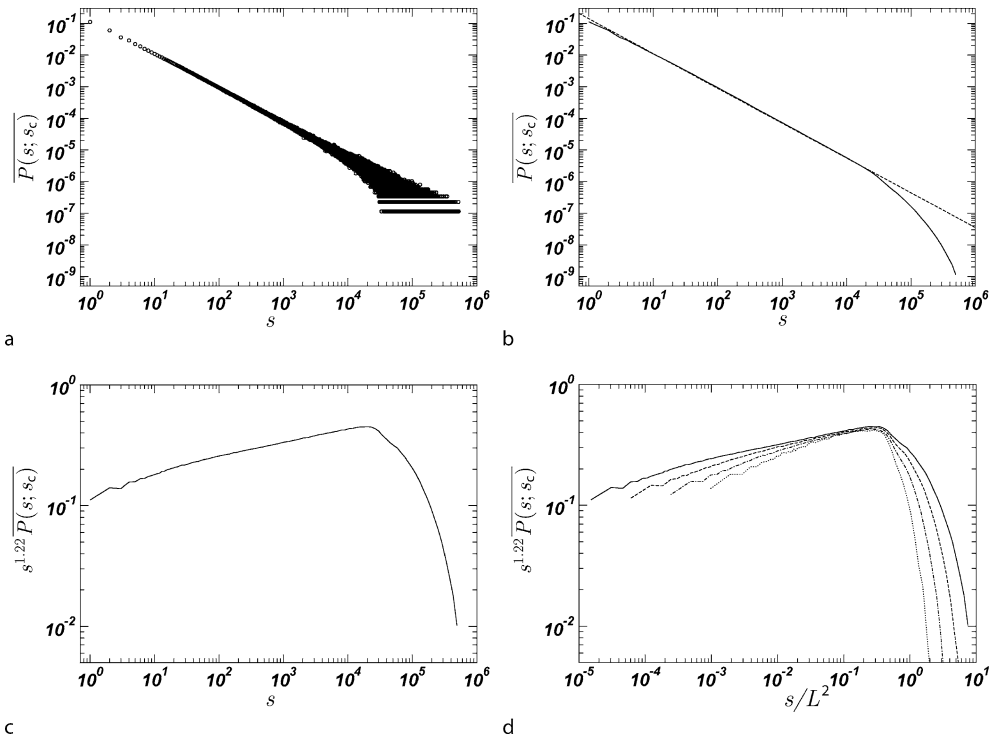
The histogram  $\overline{P(s)}$ , calculated by repeating the above procedure millions of times, measures the frequency with which a certain avalanche size has been observed. It represents a numerical estimate of the exact, discrete probability function  $\langle P(s) \rangle$ . In the following, an over-line as in  $\overline{P(s)}$  indicates a numerical estimate, while brackets, such as in  $\langle P(s) \rangle$  indicate the exact population average.

Strictly, the distribution  $\langle P(s) \rangle$  is usually inaccessible, because even a finite complex system has so many different states and therefore so many different outcomes, some with very small frequencies, that it is virtually impossible

to calculate the distribution exactly or sample all outcomes with the exact weight.

As shown in Fig. 1, the PDF of the avalanche size has a very long, power law tail. This is typical for complex systems and distinguishes it from what is often found in simpler problems. Most importantly, the PDF deviates from a simple Gaussian which is, by the central limit theorem [6], the expected PDF, if the observable was a sum or an average of many independent random variables. The lack of convergence to a Gaussian therefore indicates the presence of **correlations**, the hallmark of a complex system. *The aim of a numerical estimate and analysis of the PDF of a complex system is to characterize it quantitatively in the form of exponents, moment ratios and other features, in order to compare it to other complex systems.* To improve the numerical results, various simulation and analysis techniques are employed, many of which are aiming to minimize the amount of CPU-time spent in the simulation.

In the following, some standard simulation techniques are described. These methods have been developed since the early 1950s [2] and represent themselves a branch of



**Probability Densities in Complex Systems, Measuring, Figure 1**

BTW simulation of a system of size  $256 \times 256$ , with  $1 \cdot 10^7$  iterations used in the transient and  $2 \cdot 10^7$  iterations used for statistics: **a** Raw histogram, **b** binned histogram, (dashed line shows exponent  $-1.1$ ) **c** rescaled histogram, using  $\tau = 1.22$  and **d** (attempt) of a data collapse including very small system sizes,  $L = 32, 64, 128, 256$ . The different ways to generate, process, present and analyze the raw data are discussed in this article

numerical analysis. The data analysis is based on the **sample**, i. e. the set of **measurements** produced in numerical simulations, subject of the following section. The analysis can either focus on the PDF itself or on its moments, revealing different features of the system.

## Simulation Techniques

In general, the complex system to be studied numerically models a more complicated phenomenon observed in nature, in economy, society etc. Typically, such a model is composed of a large number of interacting elements, which might be divisible in different classes. These elements might be agents, particles, organisms, locations etc. and the classes might represent different rôles or species. Depending on the class, the type of interaction might be very different.

Only in very rare cases the model strictly reflects the original natural, economical or social observation it has been derived from. A complete representation of the original problem, even if possible, might be numerically intractable, because of too large a number of parameters that characterize the model. These parameters generally are **couplings**, i. e. they parametrize the interaction between the individual elements, and generally stay fixed during the simulation. The immediate aim of the simulation is to determine observables as functions of the parameters. Only if the number of parameters is small enough, the **parameter space** can be covered sufficiently densely to allow for a reliable statement of their various rôles and effects.

The parameter space is to be distinguished from the **phase space**, which is the space or set of possible configurations of the system. Here, a **configuration** or **state** is the (smallest) set of values describing the system in such a way that they suffice at any point to prescribe the further evolution of the system. Each of these values is associated with a **degree of freedom** in phase space.

Reducing the degrees of freedom generally has little impact on the quality of the numerical estimates. Even the simplest computer models usually have so many states that it becomes virtually impossible to calculate any property by realizing all such states. For example, a two dimensional square lattice of size  $10 \times 10$  sites, which each can be in one out of two states, has  $2^{100}$  possible states and it would take thousands of billions of years to visit each state, even if a billion states could be realized every second. Such a comprehensive study of this rather small system therefore is out of scope. Moreover, the original process to be modeled, realized in nature or society, cannot be thought of being appropriately represented in this form.

Monte Carlo is by far the most widely used technique for measuring probability densities in a complex system. Other approaches, in particular deterministic **molecular dynamics**, are far less popular. This section focuses on the practice of Monte Carlo and various standard techniques used in conjunction with it.

## Monte Carlo Methods

In its most general form, a Monte Carlo algorithm estimates an expectation value of an observable by averaging over a random sample. In the following, this is illustrated for a problem with discrete configurations  $\sigma$ , each of which describing the state of the entire system. In case of interacting particles at sites, this would be a set of numbers representing the state of every individual site. Assuming the exact, but usually unknown probability of state  $\sigma$  to be  $P(\sigma)$ , which in systems with continuous degrees of freedom is replaced by a PDF, the expectation value for an observable  $A(\sigma)$  is

$$\langle A \rangle = \sum_{\{\sigma\}} P(\sigma) A(\sigma) \quad (2)$$

where the sum runs over the set of all states  $\{\sigma\}$ . The exact expectation value can be approximated by the mean  $\bar{A}$  over a (small) **random sample** of  $N$  configurations  $\sigma_1, \sigma_2, \dots, \sigma_N$ ,

$$\bar{A} = \frac{1}{N} \sum_{i=1}^N A(\sigma_i) \quad (3)$$

if the configurations  $\sigma_i$  occur with  $P(\sigma)$ . By the (weak) law of large numbers, the estimate  $\bar{A}$  converges to the exact value  $\langle A \rangle$  in the limit of a large **sample size**  $N$ .

In most classical models of statistical mechanics, the probability  $P(\sigma)$  is known exactly and often corresponds to the Boltzmann–Gibbs weight. In some rare cases, such as percolation [7], configurations of the system are all equally important and are therefore generated easily at random and independently. For a much greater class of problems, for example in case of the Ising model [8], the configurations differ greatly in statistical weight and it is a hard computational problem to generate them in a sensible way. In these cases, Monte–Carlo sampling is mainly concerned with finding *important* configurations, which is often achieved by constructing a pseudo-dynamical process that generates the configurations with the correct probabilities. In complex systems, on the other hand, usually the dynamics is given while the probability is unknown, so that the first concern is to produce measurements with the correct probability (density). However,

both issues lead to the same solution, namely **Markov Chain Monte Carlo**, which can be regarded as a form of **importance sampling**.

Importance sampling is a method for producing measurements with a frequency optimized for reducing the number of measurements required to estimate an expectation value within a prescribed error. The sampling method therefore is, ideally, adapted to the observable. In the vast majority of numerical simulations of complex systems, however, the sampling scheme used corresponds to the simplest form of a Markov chain, where the configuration of the system evolves from one stage to another in discrete steps given by the dynamics, so that the frequency with which a configuration occurs during the simulation converges to the exact probability of that configuration [9]. The Markov condition means that the probability to observe a certain configuration of the system depends *only on the directly preceding configuration*. By including a sufficient amount of information in the configuration, in principle every process can be rendered Markovian. For example, while the sequence of particle coordinates in a classical gas is not Markovian, the same sequence together with the particles' moments is so.

More sophisticated approaches bias the sampling so that regions in phase space which have a greater contribution  $A(\sigma)P(\sigma)$  to the expectation value than other regions are visited more frequently, which is discussed further in Subsect. "Rare Events".

### Monte Carlo Applied to Complex Systems

In the following, it will be assumed that the complex system has its dynamics given in the form of a set of rules specifying how the model evolves from one configuration to another. Furthermore, it will be assumed that these rules involve some randomness, so that a given starting configuration does not necessarily lead to one particular configuration in the next step. The evolution of the system is subject to certain **transition probabilities** encoded in the transition matrix  $W(\sigma, \sigma')$ , which gives the probability to go from configuration  $\sigma'$  to configuration  $\sigma$ . This leads to the notion of the system's **trajectory in phase space**, which is a random variable, traced out by application of the transition matrix in every **Monte Carlo time step or update**. **Markov Chain Monte Carlo** uses the phase space trajectory to sample the phase space in a fair fashion.

By construction, consecutive configurations are generally correlated, i. e. the probability to find the system in a certain configuration depends on the preceding configuration. Given the Markovian nature of the process, the probability depends only on the previous configuration,

nevertheless correlations can be very long-lived, because the previous configuration's probability depends in turn on its preceding configuration and so on. The **correlation time** is a measure for the number of updates over which these correlations decay, the numerical consequences of which are discussed in Subsect. "Correlation Time".

At the beginning of the simulation the system is initialized in a particular form which in practice is often a rather artificial state. Due to correlations the influence of the initial condition can last for very many updates. The Markov-chain approach relies on the insight that the sampling frequency after a sufficiently long **transient** converges to the **stationary distribution**  $P(\sigma)$ , which solves for all states  $\sigma$  [6]

$$0 = \sum_{\{\sigma'\}} P(\sigma') W(\sigma, \sigma') - \sum_{\{\sigma'\}} P(\sigma) W(\sigma', \sigma) \quad (4)$$

where  $W(\sigma, \sigma')$  is the transition probability from  $\sigma'$  to  $\sigma$  introduced above. Unless relaxation processes are to be studied, the stationary regime is sampled after a sufficiently transient, which is ignored, i. e. configurations generated during the transient do not enter the numerical estimates. It is general practice to save the configuration of a system regularly at later stages, which might subsequently be used as a starting point for other trajectories in phase space. This method also allows a systematic study of the influence of the transient. Eq. (4) is to be distinguished from **detailed balance**,

$$0 = P(\sigma') W(\sigma, \sigma') - P(\sigma) W(\sigma', \sigma) \quad (5)$$

which solves (4) term by term. Detailed balance is the hallmark of equilibrium thermodynamics and has far reaching consequences for the topology of the network spanned by configurations and the possible transitions between them [10].

**Poisson Processes** Many complex systems consist of a set of concurrent **Poisson processes**, for example particles at sites that decay with one particular rate and interact with another rate. Maintaining a list of all possible processes and treating the fastest as deterministic is a widely adopted approach. If  $n$  processes with rate  $r_1, \dots, r_n$  are possible with the fastest having the largest rate  $e$ , one process, say  $i$ , out of  $n$  is picked at random with uniform probability, performed with probability  $r_i/e$  [11] and the continuous time incremented by  $1/(en)$ . Within time  $T$  on average  $enT$  picks are made, each process is selected  $eT$  times and performed  $r_i T$  times, which is precisely the average number of times the Poissonian event should occur. This implementation therefore correctly reproduces the averages. On the other hand, an event controlled by a Poisson

process can take place an arbitrary number of times within any finite time-span, which this implementation is incapable of. If correct time-keeping is expected to play an important rôle, as for example in the case of temporal correlations, the time increment itself is to be drawn randomly from a waiting time distribution [6,12].

**Rare Events** In the presence of rare events, [13] reliably estimating expectation values might become impossible. The problem affect bounded and unbounded observables, although an unbounded observable is potentially more prone to this phenomenon.

The problem is caused by events that occur only very rarely, where, however, the observable is particularly large. If, for example, an event occurs with frequency  $f = 10^{-6}$  where the observable  $A$  has a value of  $A^* = 10^4 a$ , with  $a$  being the average of  $A$  over all other realizations, then including the event will change the estimate of the average of  $A$  only by  $fA^*/a = 10^{-2}$ . Assuming that the second moment over all other events is  $a^2$ , including the rare event will shift the estimate by  $fA^{*2}/a^2 = 10^2$ .

An even more dramatic effect is expected for higher moments, the details of which are discussed in Sect. “Scaling”. Anticipating Subsect. “Moment Analysis” somewhat, the difficulty of estimating higher moments reliably lies in their higher demand of independent measurements (also discussed in Subsect. “Estimators”). To prevent systematic underestimation, at least a certain fraction of measurements must be taken from above  $\langle s^n \rangle^{1/n}$ . The number of measurements required increases significantly with  $n$  and is further increased in cases of long time correlation and pathological distributions. In the presence of rare events more sophisticated importance sampling techniques have to be used [14,15] to calculate reliable estimates.

In financial applications, rare events are often discussed in the context of so-called “fat tails” [16], which refer to PDFs that are close to a normal distribution but still have significant support at about five to ten standard deviations away from the mean. Fat tails are sometimes characterized by power laws, which are often the focus of the analysis of complex systems, as discussed in this article.

Rare events are sometimes associated [17] with multiscaling, which is a form of power law scaling to be discussed in Subsect. “Gap Scaling Versus Multiscaling”.

**Random Number Generators** The simulation of a complex system that is essentially discrete and can be represented by integers usually spends a significant fraction of time producing (pseudo) random numbers. If, on the other hand, the model is essentially continuous, i.e. requires floating point numbers and involves functions from

the mathematical library, those are likely to consume most of the CPU-time.

This observation suggests two paths of optimization: Firstly, floating point operations should be avoided as much as possible. For example, as most random number generators produce random integers, a frequent comparison of the form  $\text{rand}() / \text{RAND\_MAX} < q$  where  $q$  is a constant floating point number, should be replaced by  $\text{rand}() < i$ , where  $i$  is  $q$  rescaled by a factor  $\text{RAND\_MAX}$ . Similarly, Boolean random variables, true with probability 1/2, should be replaced by random bits, obtained using a bit-mask on an integer random variable, which is shifted on many platforms most efficiently by adding it to itself.

The second obvious path of optimization addresses the (pseudo) random number generator. All modern random number generators, such as `rand2` and others from [18], various generators discussed in [19] or the Mersenne twister [20], are of comparable quality and usually have passed all standard tests [21,22]. Their periods, that is the number of pseudo random numbers generated until they repeat, are usually long enough for modern requirements. Different generators mainly differ in CPU-time consumptions and general acceptance. Linear congruential random number generators and some of those found in the Standard C Library, on the other hand, generally lack the quality required in modern computer simulations. They often fail standard tests, yet some of them are very fast.

Different numerical simulations are independent only if the random number sequences used are independent. Only very few random number generators, mostly those designed for the use in parallelized simulations, guarantee the independence of sequences for different seeds, which in some cases have to be generated separately. It is good practice to seed the random number generator in a controlled way that ensures that none of the seeds is used more than once.

## Estimators

Similar to experiments, numerical simulations suffer from different sources of error, many of which can be reliably estimated or even cured. In a Monte-Carlo simulation, expectation values are estimated using an **estimator**, for example the estimator for the  $n$ th moment

$$\langle s^n \rangle = \int ds s^n P(s) \quad (6)$$

from a simulation producing a sequence  $s_1, s_2, \dots, s_N$  is

$$\overline{s^n} = N^{-1} \sum_{i=1}^N s_i^n. \quad (7)$$



For the sake of the argument, in the following it will be assumed that all moments exist.

An estimator is called **unbiased** if the expectation value of the estimator coincides with the expectation value of the object to be estimated, independent of the sample size  $N$ . Provided the sequence  $s_1, \dots, s_N$ , introduced above, was taken from the stationary regime, the above example, Eq. (7), estimating  $\overline{s^n}$ , is unbiased, because  $\langle \overline{s^n} \rangle = \langle s^n \rangle$  since  $\langle s_i^n \rangle = \langle s^n \rangle$  for every element in the sequence. On the other hand,  $\overline{s^{n^2}}$  is not an unbiased estimator of  $\langle s^n \rangle^2$ , even if the elements in the sequence are independent. In this case,

$$\langle \overline{s^{n^2}} \rangle = \langle s^n \rangle^2 + N^{-1} (\langle s^{2n} \rangle - \langle s^n \rangle^2) \quad (8)$$

which has an explicit dependence on  $N$  unless the variance of  $s^n$  vanishes. Consequently, the unbiased estimator of the variance of  $s^n$ , denoted as  $\overline{\sigma^2}(s^2)$  is

$$\overline{\sigma^2}(s^n) = \frac{N}{N-1} (\overline{s^{2n}} - \overline{s^n}^2) \quad (9)$$

assuming mutual independence of the measurements.

An estimator is called **consistent** if it converges to the quantity to be estimated as the sample size is increased. The law of large numbers ensures this property for simple means, however, more complicated objects to be estimated might not possess an estimator that is obviously consistent.

**Numerical Error** The central aim of a numerical simulation is to produce, efficiently, large sample sizes, with small or even vanishing correlation time, estimating observables using unbiased and consistent estimators. The quality of the numerical estimate, i. e. its error is usually measured by means of the standard deviation of the estimator, which, in turn, is to be estimated as well. The standard deviation measures the width of the Gaussian describing the expected normal distribution of the estimator.

The standard deviation,  $\sigma(\cdot)$ , is defined as the square root of variance  $\sigma^2(\cdot)$ . The variance of the estimator of  $\langle s^n \rangle$ , introduced in Eq. (7) is

$$\begin{aligned} \sigma^2(\overline{s^n}) &= \langle \overline{s^{n^2}} \rangle - \langle \overline{s^n} \rangle^2 = N^{-1} (\langle s^{2n} \rangle - \langle s^n \rangle^2) \\ &= N^{-1} \sigma^2(s^n) \end{aligned} \quad (10)$$

using Eq. (8), assuming that the sample is uncorrelated. A naïve estimator of this variance is  $N^{-1}(\overline{s^{2n}} - \overline{s^n}^2)$ , which is, however, a biased estimator, because of Eq. (8):

$$\left\langle N^{-1} (\overline{s^{2n}} - \overline{s^n}^2) \right\rangle = N^{-1} \sigma^2(s^n) \frac{N-1}{N} \quad (11)$$

so that the unbiased estimator for the variance of the estimator of  $\langle s^n \rangle$  is in fact

$$\overline{\sigma^2}(\overline{s^n}) = (N-1)^{-1} (\overline{s^{2n}} - \overline{s^n}^2). \quad (12)$$

When estimating functions of means, such as the variance

$$f(\langle s \rangle, \langle s^2 \rangle) = \langle s^2 \rangle - \langle s \rangle^2 \quad (13)$$

finding an unbiased estimator for it might already be a difficult task. The most natural choice is of course  $f(\overline{s}, \overline{s^2})$  but as pointed out above, Eq. (9), this choice is biased.

The task of finding an estimator for the variance of the estimator might be even more problematic. The most naïve approach is to assume independence and approximate the variance by error propagation

$$\sigma^2(f) \approx \sigma^2(\overline{s}) \left( \frac{\partial f}{\partial \overline{s}} \bigg|_{\langle \overline{s} \rangle, \langle \overline{s^2} \rangle} \right)^2 + \sigma^2(\overline{s^2}) \left( \frac{\partial f}{\partial \overline{s^2}} \bigg|_{\langle \overline{s} \rangle, \langle \overline{s^2} \rangle} \right)^2. \quad (14)$$

A first attempt to include correlations is

$$\begin{aligned} \sigma^2(f) &\approx \sigma^2(\overline{s}) \left( \frac{\partial f}{\partial \overline{s}} \right)^2 + \sigma^2(\overline{s^2}) \left( \frac{\partial f}{\partial \overline{s^2}} \right)^2 \\ &\quad + 2 \text{covar}(\overline{s}, \overline{s^2}) \frac{\partial f}{\partial \overline{s}} \frac{\partial f}{\partial \overline{s^2}} \end{aligned} \quad (15)$$

where  $\text{covar}(\overline{s}, \overline{s^2})$  denotes the covariance

$$\text{covar}(\overline{s}, \overline{s^2}) = \langle \overline{s \overline{s^2}} \rangle - \langle \overline{s} \rangle \langle \overline{s^2} \rangle. \quad (16)$$

Similar schemes are often used to estimate variances of very complicated functions of the observables, sometimes assuming that the expectation value of the function coincides with the function of the expectation values of the observables, which clearly is only the case for linear functions.

One widely used class of methods available to reduce bias in estimators and the estimators of their variances are the Bootstrap and the Jackknife [23,24]. Both methods are so-called “resampling plans”, which prescribe a method to construct estimators and estimate their variances from the distribution of estimates based on sub-samples. In case of the Jackknife, a sample containing  $N$  individual measurements produces  $N$  sub-samples by taking all measurements apart from the first, second, third and so on. The estimator of the desired quantity is applied to each of the  $N$  sub-samples, each of which containing  $N-1$  measurements. More elaborate schemes are available particularly suited to small sample sizes.

A very similar approach is to measure the desired quantity in, say  $M$ , independent measurements, such as different Monte-Carlo simulations with different random number sequences, each consisting of  $N$  individual measurements. If  $N$  is large enough, the central limit theorem renders the distribution of the  $M$  estimates indistinguishable from a normal distribution. The error of the mean is then easily derived. This scheme lends itself to a simulation consisting of many independent runs, if they are necessary anyway to determine a quantity reliably. It can be applied in a weighted fashion if individual runs consists of different sample sizes.

**Correlation Time** If the sequence  $s_1, s_2, \dots, s_N$  used in Eq. (7) is observed in a Markov process, the individual measurements will not be mutually independent: If  $P(s_2|s_1)$  denotes the probability to observe  $s_2$  given the directly preceding sample was  $s_1$ , then  $P(s_2|s_1)$  is, in general, not independent of  $s_1$ . The same is true for  $n > 1$  in  $P(s_n|s_1)$ , which is generally not independent of  $s_1$ , and even  $P(s_n|s_{n-1}, s_{n-2}, \dots, s_1)$  or any other probability of  $s_n$  given a previous (sub-)sequence. Only if the probability of  $s_n$  depends only on the directly preceding sample  $s_{n-1}$  but not on any earlier one, then the sequence  $s_1, \dots, s_n$  itself is a Markov process and fully specified by the transition probabilities between successive  $s_i$ .

The **independence** of two measurements  $s_i$  and  $s_j$  means that their joint probability factorizes. It implies that they are **uncorrelated**, which means that the expectation value of their product factorizes,  $\langle s_i s_j \rangle = \langle s_i \rangle \langle s_j \rangle$ . The converse is generally not true, but frequently assumed.

The correlation time in a sample is derived from the **autocorrelation function** in the stationary state,

$$C(j) = \frac{\langle s_i s_{j+i} \rangle - \langle s \rangle^2}{\sigma^2(s)} \quad (17)$$

which is independent of  $i$  because of stationarity. The autocorrelation function is normalized with the variance of the observable  $s$ ,

$$\sigma^2(s) = \langle s^2 \rangle - \langle s \rangle^2 \quad (18)$$

so that  $C(0) = 1$ . Motivated by the study of Markovian processes [6], the **correlation time**  $\tau$  of a sequence is estimated by fitting  $C(j)$  to an exponential  $\exp(-j/\tau)$ . One can show that the mean calculated from a finite sequence  $s_i$ ,  $i = 1, 2, \dots, N$ , with correlation time  $\tau$  has a variance equal to the variance of the mean derived from a set of  $N/(2\tau + 1)$  uncorrelated measurements.

Obviously, in numerical simulations a sample with large correlation time is inferior to one with a smaller

correlation time. The estimated mean from an uncorrelated sample comprising of  $N$  measurements has variance  $\sigma^2(\bar{s}) = \sigma^2(s)/N$ , so that along the simple arguments presented above, the variance of the mean increases (almost) linearly in the correlation time and decreases inversely in the sample size. As discussed further below, the square root of the variance of the estimator is usually used as a measure of the statistical error. When comparing different numerical techniques in terms of their computational costs, i. e. their CPU-time, one therefore has to compare the product of the square of the error and CPU-time.

**Reducing the Correlation Time** A sample of **sample size**  $N$  and correlation time  $\tau$  contains  $\tau$  correlated sub-samples which have only correlation time 1,  $s_1, s_{1+\tau}, s_{1+2\tau}, \dots$ . By pruning the original sample or, equivalently, reducing the sampling rate in a numerical simulation, one can reduce the correlation time to arbitrarily small values. This, however, is achieved only with the additional computational cost for the intermediate states of the system which are produced but not sampled. Such a technique pays off only if the cost of the sampling is high compared to the production of a new state of the system, as for example if the (expensive) Fourier transform is to be taken in a system the configuration of which evolves at virtually no costs. Often, it is more efficient to include all correlated sub-samples.

## Scaling

Scaling is the hallmark of universality [25,26]. One of the aims of the analysis of a complex system is to determine whether it displays scaling and if so, whether the characteristics of the scaling behavior suggest that it belongs to a certain universality class. The **universality hypothesis** makes it possible to extend the predictive power of a simple model to more realistic situations and entire classes of stochastic processes.

## Finite Size Scaling Hypothesis

The PDF  $\langle P(s; s_c) \rangle$  of an observable expected to display scaling is tested against the scaling hypothesis also known as **simple scaling**

$$\langle P(s; s_c) \rangle = \begin{cases} a s^{-\tau} G(s/s_c) & \text{for } s > s_0 \\ f(s; s_c) & \text{otherwise} \end{cases} \quad (19a)$$

$$(19b)$$

where  $a$  is a **metric factor** (constant and in particular independent of  $s_c$ ),  $\tau$  denotes a **(critical) scaling exponent** (not to be confused with the correlation time) and  $G(x)$  is the **scaling function** usually assumed to be universal up to

a pre-factor. The **upper cutoff**  $s_c$  is the typical scale of the observable for which the density function rapidly drops to 0. It is also called “characteristic event size” and coincides, up to a pre-factor, with the first moment  $\langle s \rangle$  in the case  $\tau = 1$ .

The **lower cutoff**  $s_0$  separates a range of values of the observable  $s$  where the histogram is governed predominantly by the non-universal PDF  $f(s; s_c)$ . Equation (19) applies only approximately in discrete systems and in general only asymptotically, i. e. for sufficiently large  $s$ ,  $s_c$ . Taking this limit appropriately, quickly becomes very technical and for that reason will not be discussed in detail in the following.

The upper cutoff  $s_c$  is often assumed to diverge with the system size  $L$  as a power-law,  $s_c = bL^D$  with  $D$  being another (critical) exponent, which can be regarded as the spatial dimension of  $s_c$  and therefore of the observable  $s$ . The pre-factor  $b$  is a second metric factor. Together with  $a$ , these two factors are the only **non-universal** parameters entering the universal part of the PDF, Eq. (19). In order to test for universal behavior, one has to impose that  $G(s/s_c)$  is a dimensionless function of a dimensionless argument. For dimensional consistency,  $a$  and  $b$  both are dimensionful, unless  $\tau = 1$  in which case  $a$  is a pure number which can in principle be absorbed into  $G$ , or if  $D = 1$ , in which case  $b$  is a number which can be absorbed into  $G$  as well.

If  $s_c$  depends only on  $L$ , (19) describes **finite size scaling**. In standard critical phenomena [25] this is observed only exactly at the critical point. Away from it, the upper cutoff depends on an external tuning parameter, such as the reduced temperature  $t$ , which vanishes at the critical point of a ferromagnetic phase transition. The PDF then is still expected to display scaling as in (19), to be investigated in the thermodynamic limit, i. e.  $L \rightarrow \infty$ . In this case one expects  $s_c = b'\xi^D$  with  $b'$  being another metric factor,  $\xi$  the correlation length and  $D$  the same spatial dimensionality as above. Similarly  $\tau$  is expected to be the same, while the scaling function differs in finite size scaling and critical (point) scaling. Most importantly, the finite size scaling function depends on more details of the system than the critical scaling function, such as the shape, topology and aspect ratio of the system as well boundary conditions [26].

The difference between finite size scaling function and critical scaling function is generally explained to be caused by the difference in correlation length: The statistics of a finite system at the critical point displaying finite size scaling depends on the geometric properties of the boundaries of the system as the correlation length reaches them. In case of critical scaling the system can often be treated as being composed of many independent sub-systems, not probing the boundaries. Therefore, the distribution of some ob-

servables, such as the order parameter, is normal if the thermodynamic limit is taken when the system is not at the critical point. Other PDFs, such as the cluster size distribution in percolation, remain non-trivial but different from finite  $L$ , as they are not subject to the central limit theorem [6,7].

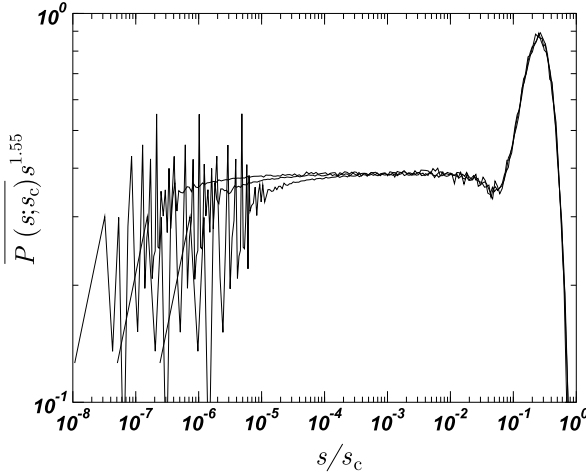
Complex systems and most notably self-organized criticality very often display finite size scaling and lack an explicit tuning parameter. Therefore, the following focuses on finite size scaling.

**Data Collapse** By plotting the numerical estimate of  $\langle P(s; s_c) \rangle(s)$  in a double logarithmic plot versus  $s$ , the slope of the resulting graph in the scaling region, i. e. the region dominated by the power-law behavior,  $s_c \gg s \gg s_0$ , gives the **apparent exponent**  $\tilde{\tau}$ . This is illustrated in Fig. 1b. According to Eq. (19a) the slope is  $-\tau$  plus any contribution due to the scaling function. Only if the scaling function converges to a constant for small arguments,  $\lim_{x \rightarrow 0} G(x) > 0$ , will the apparent exponent  $\tilde{\tau}$  coincide with the (actual) scaling exponent  $\tau$ .

From the slope of the PDF in a double-logarithmic plot a first estimate for  $\tau$  can be derived and can subsequently be used in a **data collapse**, which consists in plotting the estimates of the PDF in the rescaled form  $s^\tau P(s; s_c)$  versus  $s/s_c$ . For each individual measurement  $s_c$  is to be estimated, but the same  $\tau$  is assumed to apply to all of them. If the result follows simple scaling, then there is a value of  $s_c$  for each measurement so that all data  $s^\tau P(s; s_c)$  for  $s > s_0$  collapse onto the same curve, representing the scaling function in the form  $aG(s/s_c)$ , as exemplified by the so-called Oslo model [27] in Fig. 2. A failed data collapse is shown for small system sizes of the BTW model in Fig. 1d. The construction of a data collapse is discussed further in Subsect. “[Binning During Data Analysis](#)”.

The different measurements with  $s_c$  varying between them are usually taken from different system sizes. Plotting  $s_c$  versus the system size in a double-logarithmic plot reveals the spatial dimension or gap exponent (see Subsect. “[Gap Scaling Versus Multiscaling](#)”)  $D$ , since  $s_c = bL^D$ , possibly with corrections to scaling (see Subsect. “[Corrections to Scaling](#)”). The same strategy is applied if  $s_c$  is expected to depend on a different parameter, such as a temperature.

There is no established method for quantifying the quality of a data collapse and it therefore remains somewhat subjective. There is, however, a simple method to rule out simple scaling, for example if the tail of the scaling function, i. e. the region of large arguments, changes shape for different values of  $s_c$ , as in the BTW model at least in the case of small system sizes, Fig. 1d, or if the lower cut-



**Probability Densities in Complex Systems, Measuring, Figure 2** Example of a data collapse, here for the avalanche size distribution of the one-dimensional Oslo model, [27], driven at a single site, with system sizes  $L = 640, 1280, 2560$ . The upper cutoff is expected to scale like  $s_c \propto L^{2.25}$ . The data collapses for large values of  $s$ . For small  $s$ , the behavior is non-universal

off changes significantly with increasing  $s_c$  as in the Forest Fire Model [28]. It is particularly important to probe for this phenomenon, because the quantitative analysis based on moments of the PDF is generally unable to detect it, see Subsect. “Variance and Numerical Error of Higher Moments”.

**The Lower Cutoff** The lower cutoff  $s_0$  appears in Eq. (19) to distinguish the universal part of the PDF shaped by  $G(s/s_c)$  from the non-universal part given by  $f(s; s_c)$ . Below the lower cutoff complex and critical systems are expected to be governed by microscopic physics, i.e. the specific details of the process. In this region, the PDF might depend on additional parameters. More often however, the processes on a very small scale are expected to be similar, so that often  $f(s; s_c) \approx f(s; s'_c)$  even for  $s_c$  very different from  $s'_c$ . This is found, for example, in critical percolation, where  $s_c$  is solely determined by the system size and  $s$  is chosen to be so small that the finiteness of the system is irrelevant. This feature is also visible in the patterns at low  $s$  in Fig. 2. Different definitions of the observable usually have an impact in this region, but not for the asymptotic, large  $s$  behavior.

Some complex systems do not possess a lower cutoff, so that the entire region of accessible values of the observable  $s$  is governed by the universal behavior. Although discrete systems have a natural lower cutoff given by the smallest possible event, this might be hidden in the defi-

nition range of the observable, as for example in the case of the avalanche size distribution of the one-dimensional BTW model, which follows  $\langle P(s) \rangle = s^{-1}(s/s_c)\theta(1 - s/s_c)$  with  $s \in \{1, 2, \dots, s_c\}$ ,  $s_c = L$  and  $\theta$  denoting the Heaviside step function.

A continuous system without a lower cutoff is physically equivalent to one without an upper cutoff, because by suitable rescaling the range of observables  $s$  tested can be made arbitrarily large.

**Corrections to Scaling** Equation (19a) describes only the leading order behavior of the PDF, which is only asymptotically valid as  $s_c \gg s_0$  and  $s \gg s_0$ . The former condition,  $s_c \gg s_0$ , suppresses contributions from the non-universal part and can in case of finite size scaling be achieved by increasing the system size. The latter condition,  $s \gg s_0$ , is reflected in the condition  $s > s_0$ , however the larger  $s_0$  is chosen, the smaller the error of Eq. (19a).

The corrections accounting for the failure of the exact scaling behavior are known as **corrections to scaling** [29], originally introduced in the context of ferromagnetic phase transitions. In principle, the scaling form Eq. (19a) can be generalized to

$$as^{-\tau}G(s/s_c) + a_1s^{-(\tau+\omega_1)}G_1(s/s_c) + \dots \quad (20)$$

with  $\omega_1 > 0$ . It is much more common to account for corrections to scaling on the level of individual observables, such as moments or  $s_c$ , instead of the entire distribution. For example, the standard finite size scaling ansatz including corrections to scaling is

$$s_c = bL^D(1 + b_1L^{-\omega'_1} + b_2L^{-\omega'_2} + \dots) \quad (21)$$

with the exponents  $\omega_i$  and  $\omega'_i$  expected to be universal.

## Moments

The data collapse described in the previous section suffers from the problem that its quality is not easily quantifiable. Only the obvious failure gives a strong indication of the absence of finite size scaling in the form Eq. (19).

A more quantitative tool is the **moment analysis**. In principle, moments can be derived from the histogram after the simulation, but the error introduced by binning methods, which are almost always a necessity, or rounding errors in case of continuous event sizes can be very big. They are easily avoided by calculating the moments during the simulation.

**Numerics of Moments** Moments are usually calculated during the simulation and provide access to the scaling be-

havior of the system without large memory or computational requirements. However, as individual contributions to moments can vary greatly in value, the quality of the estimate relies crucially on the precision of the numerical calculation.

The standard method to calculate the  $n$ th moment of  $P(s; s_c)$  is to sum all  $n$ th powers of individual event sizes  $s$  and finally divide by the number of contributions. If the process takes place in continuous time, each event size might carry an additional weight with it, and the normalization required is the sum over all weights. This weight usually corresponds to the amount of time the observable had a particular value now entering the estimator.

Depending on the type of process, measuring the moments can contribute significantly to the overall computing time. One method to minimize it, is to reduce the number of additions and multiplications. An implementation could read

```
// Normalization
power=weight;
moment[0]+=power;
// First moment
power*=event_size;
moment[1]+=power;
// Second moment, etc.
power*=event_size;
moment[2]+=power;
power*=event_size;
moment[3]+=power;
...
```

which can be further simplified in the form of a loop. Moments of very high order can be generated in the form

```
// Normalization
moment[0]+=weight;
// First moment
power=event_size;
moment[1]+=power*weight;
// Second moment
power*=power;
moment[2]+=power*weight;
// Fourth moment
power*=power;
moment[3]+=power*weight;
...
```

Precision is crucial in particular when it is a priori unknown whether the main contribution to a moment is due to many small events or a few very big ones. Integer variables are an option only if the weight is integer-valued or can be rendered so. They are ideal, because they do not suf-

fer from rounding errors. To avoid overruns, many platforms offer 64 bit integers. Integers are often computationally, i. e. in terms of CPU-time, advantageous as well.

Where floating point numbers are a necessity, they have to be of appropriate size. A central criterion is whether small events can still enter with sufficient accuracy at the end of the simulation, when the various variables holding *sums* of different powers of the observables contain very large numbers. Adding a sufficiently small number to a large floating point number might not actually change it, depending on the size of the mantissa. The IEEE 754 standard [30] describes floating point numbers with 24, 53 and 64 bit mantissae.

It is computationally very inefficient to study moments  $\langle s^n \rangle$  for  $n \notin \mathbb{N}$ , because it requires a floating point operation, such as `sqrt` or `pow`. Where such moments are unavoidable, these operations can often be “recycled”, by, say, calculating `pow(event_size, 1./3.)`, taking its square and constructing from these two values plus `event_size`, all powers  $1/3, 2/3, 1, 4/3, 5/3, \dots$  using a minimal number of multiplications.

**Moment Analysis** Simple finite size scaling, Eq. (19), implies that the moments  $\langle s^n \rangle$  of the PDF scale like a power of the upper cutoff  $s_c$ ,

$$\lim_{s_c \rightarrow \infty} \frac{\langle s^n \rangle}{s_c^{n+1-\tau}} = a \lim_{y \rightarrow 0} g_n(y) \quad \text{for } n > \tau - 1 \quad (22)$$

where the **amplitude**  $g_n(0)$  is a moment of the universal scaling function

$$g_n(y) = \int_y^\infty dx x^{n-\tau} G(x), \quad (23)$$

which is universal up to a pre-factor, i. e. any ratio of these amplitudes is universal. Assuming that the scaling function  $G(x)$  is continuous and has no singularity at finite argument, one can prove that the limit  $\lim_{y \rightarrow 0} g_n(y)$  exists. Moreover, as  $g_0(0) > 0$ , the exponent  $\tau$  cannot be less than unity.

Simplifying Eq. (22)

$$\langle s^n \rangle \propto s_c^{\gamma_n} \quad \text{with } \gamma_n = 1 + n - \tau \quad \text{for } n > \tau - 1 \quad (24)$$

the scaling exponents  $\gamma_n$  are usually determined for  $n$  ranging from 1 up to typically 8 as the slope in a double-logarithmic plot of the moment versus the system parameter, which in case of finite size scaling is the linear size of the system. If no such parameter is quantifiable, moments can be measured with respect to each other, noting that

$$\langle s^n \rangle \propto \langle s^m \rangle^{\frac{\gamma_n}{\gamma_m}} \quad \text{for } n > \tau - 1. \quad (25)$$



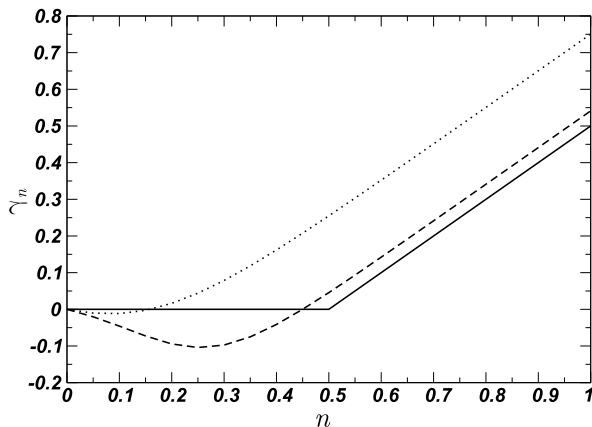
A priori the lower cutoff and the corrections to scaling are unknown and in order to determine the (asymptotic) finite size scaling behavior, a numerical study has to probe system sizes as large as possible. The true asymptote will, by definition, strictly remain inaccessible.

**Gap Scaling Versus Multiscaling** Moments following the scaling behavior prescribed by (22) are said to display **gap scaling** with  $D$  being the **gap exponent** [31]. This term refers to the constant gap between the scaling exponents for consecutive moments as a function not of  $s_c$  itself but of the tuning parameter, usually the system size. Assuming Eq. (21) and using the simplified notation Eq. (24) finite size scaling implies

$$\langle s^n \rangle \propto L^{\gamma'_n} \quad \text{with} \quad \gamma'_n = D(1 + n - \tau) \quad (26)$$

so that two consecutive moments have an exponent differing by  $\gamma'_{n+1} - \gamma'_n = D$ , visible as a constant slope when plotting the scaling exponent as a function of  $n$ , see Fig. 3. Measuring this slope is the standard method for calculating the exponent  $D$ .

The intersection of the linear continuation of the  $\gamma_n$  with the abscissa gives  $\tau - 1$ , which is a standard method of estimating  $\tau$ . An exponent  $\gamma_n > 0$  indicates that the corresponding moment is diverging as  $s_c$  diverges. For  $\tau > 1$  there are some moments for non-negative  $n$  which do not diverge. If  $\tau = 1$  the only non-negative moment that does not diverge is the normalization  $n = 0$ .



**Probability Densities in Complex Systems, Measuring, Figure 3** The exponents  $\gamma_n$  estimated from the moments calculated from  $\langle P(s; s_c) \rangle$  as introduced in Fig. 4. The dotted line shows the exponents estimated from  $s_c$  ranging from 25 to  $1 \cdot 10^4$ , the dashed line shows exponents based on  $s_c$  ranging from  $1 \cdot 10^4$  to  $6.4 \cdot 10^5$ . The full line shows the exact values the numerical data converges to. The rounding in the numerical estimates can be mistaken as multiscaling

If the exponents  $\gamma_n$  do not increase linearly in  $n$  yet the moments still display scaling behavior, the system is said to exhibit multiscaling. In this case, the scaling form Eq. (22) is often extended to

$$\lim_{s_c \rightarrow \infty} \frac{\langle s^n \rangle}{\ln(s_c)^{\lambda_n} s_c^{\gamma_n}} \quad (27)$$

to allow for logarithmic contributions to the scaling behavior. The presence of these logarithmic “corrections” are often interpreted as a sign of multiscaling, as is the presence of **rare events**, see Subject. “Rare Events”.

Multiscaling often means that the  $\gamma_n$  converge for large  $n$  and only the  $\lambda_n$  continue to increase. Multiscaling is sometimes confused with the absence of scaling altogether, or with the rounding effect observed close to  $n = \tau - 1$ , as shown in Fig. 3.

The standard method of estimating  $\tau$  and the gap exponent  $D$  is to fit the moments  $\langle s^n \rangle$  to a power-law, see (24) and Eq. (26), possibly including corrections to scaling (Subject. “Corrections to Scaling”) and usually by plotting not versus  $s_c$  but versus some system parameter such as the size  $L$ . The resulting  $\gamma_n$  or  $\gamma'_n$  are then fitted against  $1 + n - \tau$  or  $D(1 + n - \tau)$  respectively to determine  $\tau$  and possibly  $D$ .

#### Variance and Numerical Error of Higher Moments

The variance of the  $n$ th moment is  $\langle s^{2n} \rangle - \langle s^n \rangle^2 = \langle (s^n - \langle s^n \rangle)^2 \rangle$ . This difference is always non-negative and comparing the scaling exponents for both contributions from Eq. (24) confirms  $\gamma_{2n} \geq 2\gamma_n$ , where the equality holds only if  $\tau = 1$ . Thus, unless  $\tau = 1$ , the standard deviation of the estimator of the  $n$ th moment is expected to scale with exponent  $(1/2)\gamma_{2n} = nD + (1/2)(1 - \tau)$ . The relative variance, also known as the relative fluctuations, therefore is expected to scale as

$$\frac{\langle s^{2n} \rangle}{\langle s^n \rangle^2} \propto s_c^{\gamma_{2n} - 2\gamma_n} = s_c^{\tau - 1} \quad (28)$$

and therefore diverges asymptotically for  $\tau > 1$  as the upper cutoff increases.

For  $\tau = 1$  the relative variance might not change with  $s_c$ , although the ratio  $\sqrt{\langle s^{2n} \rangle} / \langle s^n \rangle$  might have a strong dependence on  $n$ . Because (24) determines only the leading order scaling of the moments but makes no statement about the respective amplitude, the difference  $\langle s^{2n} \rangle - \langle s^n \rangle^2$  might have a leading order that scales slower than  $\gamma_{2n}$  if  $\tau = 1$ . In this case the relative variance might asymptotically vanish, which is known as **self-averaging** [32].

Self-averaging is more commonplace in the context of scaling with system size away from a critical point. In this

case, spatial densities usually follow a Gaussian distribution and the system can be decomposed into finite patches, the number of which grows linearly in the volume. Therefore, the relative variance of a density that converges in the thermodynamic limit to a finite value, decreases like  $L^{-d}$ . This effect is known as **strong self-averaging** [33]. Where the variance decays with increasing  $L$ , but not as fast as  $L^{-d}$ , the effect is known as **weak self-averaging**. The variance can decrease faster than  $L^{-d}$  only in the (rare) presence of anti-correlations.

A Monte–Carlo simulation can be regarded as a means to integrate a function, so that estimating  $\langle s^n \rangle$  means in fact to perform the integral

$$\overline{s^n} = \int ds \overline{P(s; s_c)} s^n. \quad (29)$$

The most efficient method to estimate  $\langle s^n \rangle$  samples exactly with a density  $\propto \langle P(s; s_c) \rangle s^n$ , which in principle can be achieved using (ideal) importance sampling. In practice, this is rarely possible and only very crude approximations to  $\langle P(s; s_c) \rangle s^n$  are available as sampling frequencies. In the context of complex systems,  $\langle P(s; s_c) \rangle$  itself is the sampling rate, in the vast majority of problems.

The comparison to ideal importance sampling suggests an alternative, much stronger criterion to assess the quality of a numerical estimate. The product  $\langle P(s; s_c) \rangle s^n$  usually has its maximum at some  $s_{\max}$  close to  $s_c$  and one might therefore compare the density of measurements around  $s_{\max}$  to the weight this range of  $s$  enters the integral (29) [34].

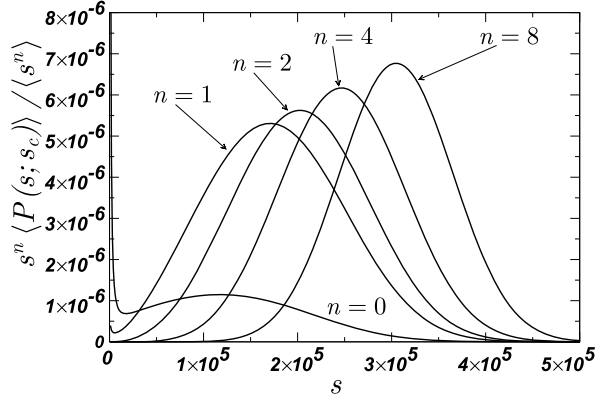
Correspondingly, one can derive the number of measurements needed to perform the integral (29) reliably by imposing a lower bound on the density of measurements around  $s$  as a fraction of  $\langle P(s; s_c) \rangle s^n$ . Demanding that the density of measurements around  $s$  is nowhere less than  $\rho^* \overline{P(s; s_c)} s^n$ , then gives rise to a scale  $s^*$ , so that at least one sample is to be taken above the “characteristic largest event”  $s^*$ ,

$$\int_{s^*}^{\infty} ds \rho^* \overline{P(s; s_c)} s^n = 1 \quad (30)$$

which depends on  $n$  and is to be compared to the total weight in the sample actually produced in the simulation  $\tilde{P}(s; s_c)$ ,

$$\tilde{N}(s^*) = N \int_{s^*}^{\infty} ds \tilde{P}(s; s_c) \quad (31)$$

where  $N$  is the total number of measurements taken and  $\tilde{P}(s; s_c) = \overline{P(s; s_c)}$  is the sampling frequency unless importance sampling is used. The event size  $s^*$  usually is of the



**Probability Densities in Complex Systems, Measuring, Figure 4** Moments for larger  $n$  draw more weight from larger  $s$ . A moment analysis investigating large moments might possibly miss features of the distribution at small arguments. The distribution used here is  $\langle P(s; s_c) \rangle = a s^{-3/2} G(s/s_c)$  for  $s \in [1, \infty$  with  $G(x) = (1 + x^2) \exp(-x^3/s_c)$  and  $s_c = 10000$

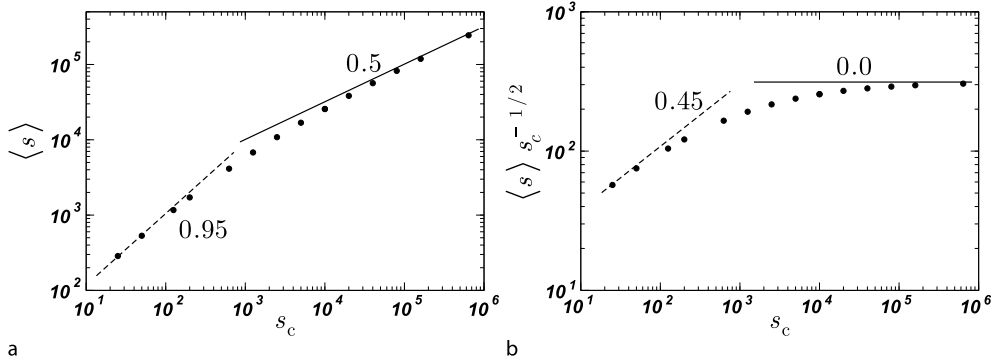
order of the upper cutoff  $s_c$  and the number of measurements needed typically scales itself like  $s_c^n$ .

Thus, this much stronger criterion, constraining the sampling density even in the extreme region of very large  $s$  produces a much stronger constraint on the sample size than suggested by the relative variance, which increases with  $s_c$  only as  $s_c^{\tau-1}$ , see Eq. (28).

Figure 4 shows the product  $s^n \langle P(s; s_c) \rangle / \langle s^n \rangle$  for a range of  $n$  as a function of  $s$ . As indicated above, the larger  $n$ , the more weight enters from the tail of the distribution, which implies that small  $s$  features are more and more ignored with increasing  $n$ . While increasing  $n$  are needed in order to derive the **gap exponent** (see Subsect. “Gap Scaling Versus Multiscaling”) reliably, an analysis that relies solely on moments can miss a lack of finite size scaling for small  $s$ , such as divergent lower cutoff.

This problem can be avoided either by careful inspection of the data collapse, or by accurate determination of the scaling of moments for very small  $n > \tau - 1$ , which might necessitate the investigation of fractional moments, see Fig. 3. The closer the moment is to  $\tau - 1$ , the slower the convergence to the asymptotic behavior. This is illustrated in Fig. 5a which shows the scaling of the first moment  $\langle s \rangle$  in a system with  $\tau = 1.5$ .

Whether displaying and analyzing the histogram or moments, it is generally significantly more informative to plot the data with the leading order divided out, for example  $\bar{s} s_c^{-1/2}$  if  $\bar{s} \propto s_c^{1/2}$  is expected as shown in Fig. 5b. If the numerical data converges to a constant, this confirms the scaling, while the shape of  $\bar{s} s_c^{-1/2}$  indicates the rôle of corrections to scaling. Moreover, the spread of the numerical



Probability Densities in Complex Systems, Measuring, Figure 5

Scaling of the first moment  $\langle s \rangle$  of  $\langle P(s; s_c) \rangle$  (see Fig. 4). a In a double-logarithmic plot, the slope ranges from 0.95 (dashed line) to about 0.5 (asymptotically exact; full line) as indicated. Correction terms (corrections to scaling) or very large system sizes are necessary for the estimated exponent to be sufficiently close to the asymptote, which is  $\gamma_1 = 0.5$  known from construction. This becomes more apparent when plotted in the form shown in b: The same data but plotted as  $\langle s \rangle s_c^{-1/2}$ . In this form it is easier to see that the asymptotic behavior is only reached by probing system sizes at least as big as the largest one shown

data across the ordinate is reduced, allowing for a more detailed inspection of the details. This is discussed further in Subsect. “Binning During Data Analysis” for the scaling of histograms.

**Critical Slowing Down** Large upper cutoffs are needed to isolate the asymptotic behavior on the one hand, but they are numerically very demanding on the other hand. Another effect which reduces the effectiveness of Monte-Carlo simulations, known from ferromagnetic phase transitions, is **critical slowing down**.

The (auto-)correlation time discussed in Subsect. “Correlation Time” effectively reduces the number of independent measurements produced in the simulation. Critical slowing down describes the behavior of the correlation time at the critical point or in finite size scaling: The correlation time  $\tau$ , see Subsect. “Correlation Time”, diverges like a power-law of the tuning parameter, which in finite size scaling means that  $\tau \propto L^z$ , where  $z$  is known as the dynamical exponent.

Taking all different effects into account, the increasing relative variance for  $\tau > 1$  (exponent  $\tau$ , see Eq. (19)), the minimal number of measurements required for a certain minimal coverage and finally critical slowing down, large system sizes always represent a challenge to the quality of the estimate of moments, in particular for large  $n$ . These are, however, necessary for a reliable estimate of the asymptotic behavior of the PDF.

**Moment Ratios** A priori the amplitude  $a$  on the RHS of (22) is unknown; definition Eq. (19a) fixes only  $\tau$  and otherwise states only that there exist quantities  $a$  and  $G(x)$ ,

fixed only up to a pre-factor, so that the universal part of the PDF obeys Eq. (19a).

By taking ratios of quantities containing non-universal pre-factors these are removed. For example,

$$\frac{g_n(0) g_1^{n-2}(0)}{g_2^{n-1}(0)} = \frac{\langle s^n \rangle \langle s \rangle^{n-2}}{\langle s^2 \rangle^{n-1}} \quad (32)$$

contains only moments of the scaling function with any non-universal pre-factor removed. Moreover, one can show that (32) converges to a non-zero value if  $\langle s^n \rangle$  follows Eq. (24).

For  $\tau = 1$  the metric-factor  $a$  becomes dimensionless and can therefore be fully absorbed into the scaling function. In ferromagnetic critical phenomena and related phase transitions, such as percolation, where  $\tau = 1$ , ratios of the form  $\langle s^{nm} \rangle / \langle s^m \rangle^n$  are universal already.

### Histogram Data Representation

Naïvely, collecting the histogram in a simulation of a complex system with integer valued event sizes amounts only to incrementing a variable `histogram[event_size]++`. While this procedure is very efficient, provided that the number of processor cycles needed to access `histogram[event_size]` is minimal, the memory requirements quickly exceed that of standard computers as  $s_c$  increases. Moreover, histogram entries for small event sizes might overflow, necessitating very large data types to accommodate the largest entry in the histogram. At the same time, entries for very large event sizes are very sparse. Using this technique for continuous event sizes means that they have to be rounded.

Various methods of **binning** are established to handle these problems. Binning is usually also used at the stage of data analysis, i.e. in the preparation of data after the actual simulation has taken place. The key is to map the complete range of event sizes to the range available in the representation of the histogram. In the context of computing, the map is effectively a hash-function, in the following  $h(s)$ . Assuming that the event size lies within a certain, finite range and is non-negative,  $s \in [0, s_{\max}] = \Omega$ , not necessarily discrete, sets  $S_l \subset \Omega$  are defined so that all  $s \in S_l$  are equivalent under the hash algorithm and distinct otherwise, i.e.  $s, s' \in S_l$  implies  $h(s) = h(s') = l$  and vice versa. The sets are usually continuous, so that the hash algorithm  $h(s)$  can be represented by a set of ranges, for example  $h(s) = l$  for  $s_l \leq s < s_{l+1}$ . The hash function is then used in the form `histogram[h(event_size)]++`.

After collecting the data into bins, they are normalized by the bin size and often shown with respect to the geometric mean of the bin ranges  $\sqrt{s_l s_{l+1}}$ . There is obviously some freedom of choice, but if the choice makes a qualitative or quantitative difference, this implies that the binning is too coarse.

### Binning Schemes Within the Simulation

Binning can take place either during the simulation or in the data analysis. In general, it is advisable to keep the simulation as simple as possible: If the binning scheme within the simulation fails, all data might be lost or rendered useless. If it fails at the stage of data analysis, it can easily be fixed and repeated. However, due to memory constraints, it is very often necessary to use binning within the simulation. In these cases, it is vital to choose the most efficient binning scheme to avoid a negative impact on the CPU time.

**Power Law Binning** The aim of binning generally is to reduce the histogram's memory requirements, while avoiding potential overflows. The first aim can be achieved by choosing the size of the ranges  $s_{l+1} - s_l$  according to the (expected) PDF, which a priori is, however, unknown. Assuming a pure power law,  $\langle P(s; s_c) \rangle = s^{-\tau}$ , the ranges would need to be  $s_l = (c(M - l))^{1/(1-\tau)}$  for  $\tau > 1$ , where  $c$  determines the size of the range and  $M$  is the maximum  $l$ . In case of discrete event sizes, obviously the ranges need to be rounded.

Even distribution of the events across the histogram has the additional advantage of equal error for each bin, provided that the error is only a function of the number of events. This might not be the case for very small events,

where correlations might be significantly larger than for larger event sizes.

Although the binning ranges described above would produce ideal bins, it is generally not advisable to use an expected result, such as the exponent  $\tau$ , in the process of collecting the raw data to measure this quantity.

**Exponential Binning** The scheme most suitable for presenting the data in a double-logarithmic plot is exponential binning. In this case the range limits are powers of some base  $s_b$ , so that  $s_l = r s_b^l$  with some pre-factor  $r$ . As a result, the data points are evenly spread with a spacing of  $\ln(s_b)$  after taking the logarithm of the abscissa. The error, on the other hand, increases towards larger  $s$  because of the smaller number of events collected in bins for larger event sizes. Approximating the PDF by a pure power law, the number of events in the bin with hash  $l$  is, up to the normalization

$$\int_{s_l}^{s_{l+1}} ds s^{-\tau} = \frac{(r s_b^l)^{1-\tau}}{\tau - 1} (1 - s_b^{1-\tau}) \quad (33)$$

for  $\tau > 1$  which decreases with increasing  $l$ . The width of each bin is  $s_{l+1} - s_l = s_l(s_b - 1)$ .

Exponential binning is frequently used and the one most suited for a direct calculation of the hash value during the simulation. Two simple methods can be distinguished for determining the hash value: Firstly, a function  $\tilde{h}(s)$  can be devised so that, for example, its integer part gives the hash value. In case of exponential binning, this function is  $\tilde{h}(s) = \ln(s/r) / \ln(s_b)$ . Using functions from the mathematical library, such as `log` is, however, computationally very expensive.

Secondly, one can compare  $s$  to the various  $s_l$  until  $l$  is found so that  $s_l \leq s < s_{l+1}$ . Since small events are the most frequent ones, this might be implemented by simply linearly increasing  $l$ . A crude estimate of the expected number of comparisons shows that this number generally quickly converges with increasing  $s_c$ , as  $\langle \ln(s) \rangle$  is finite.

The constant number of expected comparisons it to be compared to the obvious "divide and conquer" or tree-search approach: Given  $M$  bins,  $s$  is compared to  $s_{M/2}$ , in the next step, depending on the outcome of the previous comparison, to  $s_{M/4}$  or  $s_{3M/4}$  etc. This method can be improved further by choosing the values to compare to so that the probability for  $s$  to be above or below are roughly equal, again assuming a certain, simplified form of  $\langle P(s; s_c) \rangle$ . The expected number of comparisons in this scheme is of the order of  $\ln(M)$ . Usually, the (logarithmic) bin size  $s_b$  is fixed, so that  $M$  increases with  $s_c$  like  $\ln(s_c)$ , i.e. the expected number of comparisons is proportional to the double logarithm of  $s_c$ . Even though asymptotically

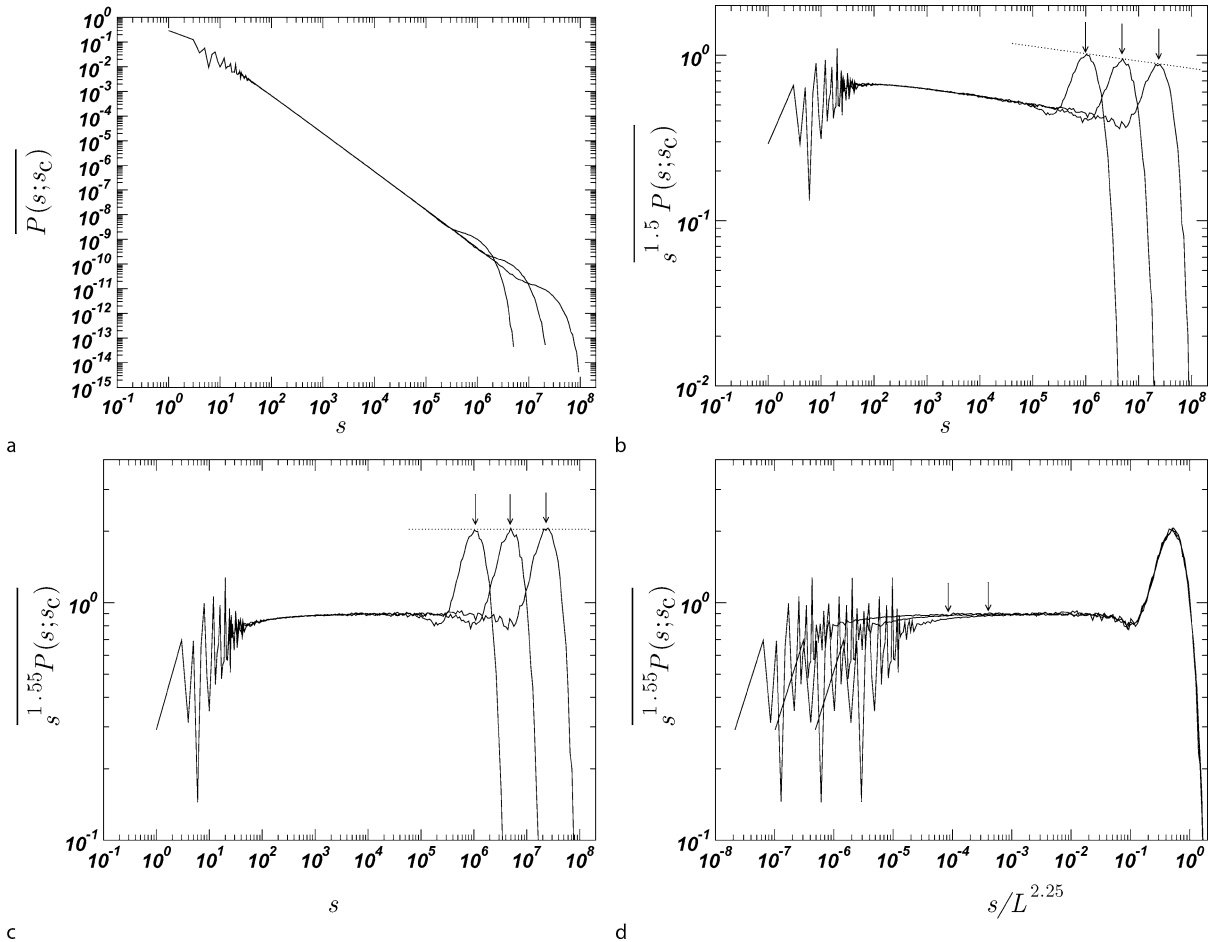
tree-search has a greater number of comparisons than the linear search discussed above, in practice, the double logarithms means that tree-search will almost always by far outperform a linear search. Similarly, even though direct calculation of the bin  $l$  using a function  $\hat{h}(s)$  has constant computational costs, they will typically be greater than for the tree search, in particular if the observable is discrete and provisions must be made in the region of its small values.

**Coarse Graining** The simplest binning scheme consists of dividing  $s$  by a constant,  $h(s) = s/q$  so that the resulting range of hash values is small enough to fit in the memory available. In case of integer valued event sizes, the simplest

and potentially the fastest way of determining the hash value is a bit shift, i. e. a division by a power of 2.

The key problem is the same as the initial motivation for binning, namely that entries for small  $s$  might overflow, while entries for large  $s$  might be very sparse. The problem is more serious in case of coarse graining if the factor  $q$  is so large that the majority of events arrives in the first bin. Moreover, not fully resolving the distribution for small  $s$  means that most of the non-universal part of the distribution is lost, potentially hiding problems of the behavior of the lower cutoff.

A very powerful alternative is a combination of coarse graining and exponential binning. A small number of thresholds is introduced and within each pair, a different



Probability Densities in Complex Systems, Measuring, Figure 6

A data collapse in detail. Using the data shown in Fig. 2, a shows the binned data. In b a preliminary estimate has been made for the scaling exponent  $\tau$  and the data is plotted in the form  $s^{1.5} P(s; s_c)$ . A "landmark" has been chosen (the maxima, indicated by arrows), which connected by a line (dotted) indicate that the exponent  $\tau$  is to be chosen larger than the preliminary choice. In c  $\tau = 1.55$  is chosen to make all maxima line up (dotted line). The position of the maxima on the abscissa (arrows) indicate the value of  $s_c$ . d shows the final collapse, after shifting the data horizontally by dividing  $s$  by the respective  $s_c$ . The approximate value for  $s_0$ , estimated for the two smaller system sizes as the point where the data approximately coincide is indicated by an arrow



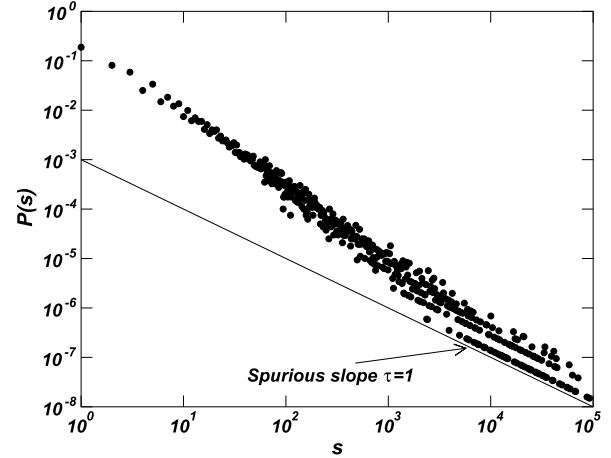
rescaling  $q$  factor is used to implement different degrees of coarse graining. This method also allows the use of different variables types in different regions of the histogram, i. e. large types for small  $s$  and smaller less memory consuming ones for smaller values.

### Binning During Data Analysis

To simplify visual inspection, in practice the raw histogram produced in the simulation is always binned afterwards. The most widely used method is exponential binning, as it provides equally spaced data points in a double-logarithmic plot.

To perform a data collapse, the data is plotted in the form  $s^\tau P(s; s_c)$  versus  $s/s_c$ . If Eq. (19) applies, the data for different  $s_c$  collapses on the same line, approximating  $G(s/s_c)$  up to an error due to **finite size corrections** and the presence of the non-universal part. An example is shown in Fig. 2. If Eq. (19) applies, the data collapses for any function  $f(x)$  when plotted in the form  $s^\tau f(s/s_c)P(s; s_c)$  versus  $s/s_c$  and shows  $G(s/s_c)f(s/s_c)$ . A particularly important case is  $f(x) = x^{-\tau}$ , which leads to  $s_c^\tau P(s; s_c)$  versus  $s/s_c$ . To expose as many details as possible of the assumed scaling function, it is generally advisable to choose  $f(x)$  so that the ordinate of the resulting data spreads as little as possible (see also Fig. 5). If  $\lim_{x \rightarrow 0} G(x) \neq 0$ , this is usually the case for  $f(x) = 1$ . If  $G(x)$  is expected to have a power-law dependence on the argument,  $G(x) = x^\alpha \tilde{G}(x)$  with  $\lim_{x \rightarrow 0} \tilde{G}(x) \neq 0$ , then  $f(x) = x^{-\alpha}$  is the appropriate choice.

If the exponent  $\tau$  and the values  $s_c$  are unknown, the first step to perform a data collapse is to collapse the binned data (Fig. 6a) approximately with some preliminary choices for  $\tau$  and (Fig. 6b) possibly also for the different values of  $s_c$  and determine an outstanding feature in the resulting data, such as the maximum (arrows in Fig. 6b). A different choice for  $\tau$  will change the relative vertical position of the “landmark”, which is to be chosen so that all landmarks line up at the same value on the ordinate, Fig. 6c. Next, the  $s_c$  for the different simulations are chosen to collapse the data horizontally, usually by estimating  $s_c$  as the position of the landmark with respect to the abscissa, Fig. 6d. In this last figure the lower cutoff  $s_0$  can be estimated as well, as the value of  $s$  from where on the data roughly coincide. A double-logarithmic plot of both cutoffs indicates roughly constant  $s_0$  and a gap-exponent  $D = 2.25$  for  $s_c$ . A data collapse often reveals that much greater system sizes are needed, as scaling applies to the *asymptotic* behavior only. The attempt of a data collapse for the BTW model, Fig. 1d, shows an example for a failed collapse in the case of small system sizes.



**Probability Densities in Complex Systems, Measuring, Figure 7**  
If the sample is too sparse in the tail of the distribution, exponential binning produces a spurious slope  $\tau = 1$

While binning greatly improves the visual quality of the data, it assumes that  $(P(s; s_c))$  does not change too suddenly on the scale of the size of the bin, which therefore needs to be chosen small enough. However, choosing  $s_b$  too small results in too many bins and therefore too much statistical noise. For large  $s$  most of those will be empty, some will contain a single entry, some two etc. Because the width of the bin is  $(s_b - 1)s_l$  the density within bins containing a single entry scales like  $s_l^{-1}$ , leading in a double-logarithmic plot to a sequence of points with slope  $-1$ . Parallel to this line, lies a set of points corresponding to bins with two entries etc. The resulting plot displays a spurious scaling exponent  $\tau = 1$ , see Fig. 7.

### Future Directions

The numerical methods discussed above are commonly used in other areas of statistical physics and have been mainly developed there. That applies to the Monte Carlo methods as well as to the data analysis. Within complexity, the development of methods hardly forms a research branch in its own right. Most methods are developed by practitioners for the specific problem at hand.

As complexity is a very diverse field, only very few models have received so much attention that specialized algorithms have been developed and discussed broadly in the literature, for example [11,35]. Most models are defined by a dynamic process which normally makes a simple implementation readily available. Algorithmic improvements mainly focus on efficiency and resources, as both CPU time and memory requirements are the two key factors limiting the quality of results in terms of sample size

and system size respectively. In other areas, as for example in the context of networks, new observables often require new, efficient algorithms.

Even though almost all models are constrained by the computer hardware available and therefore benefit significantly from improved algorithms, optimization often clashes with the universality of the algorithm and, on a more technical level, with the elegance and readability of the code.

With respect to data analysis, some literature is concerned with new approaches to established observables, such as the moment analysis discussed in Subsect. “Moment Analysis” to derive the exponents of the PDF.

While algorithms and methods have advanced significantly, there seems to be a maximum amount of information about a complex system accessible through a given amount of CPU time spent in a computer simulation. The most significant general advancement in computational physics therefore comes from the continuous improvement of computer hardware available, producing an ever increasing sample size within a certain amount of CPU time. While this relieves the research from many constraints, it does not imply that sophisticated algorithms and well thought-out methods become less relevant. It is the combined effect of methods, software and hardware that moves the subject forward continuously. A good algorithm to cope with today’s technical constraints will do an even better job in the years to come.

## Bibliography

### Primary Literature

- Metropolis N, Ulam S (1949) The Monte Carlo Method. *J Am Stat Ass* 44:335–341
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of State Calculations by Fast Computing Machines. *J Chem Phys* 21:1087–1092
- Landau DP, Binder K (2000) *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, Cambridge
- Bak P, Tang C, Wiesenfeld K (1987) Self-Organized Criticality: An Explanation of  $1/f$  Noise. *Phys Rev Lett* 59:381–384
- Bak P, Tang C, Wiesenfeld K (1988) Self-Organized Criticality. *Phys Rev A* 38:364–374
- van Kampen NG (1992) *Stochastic Processes in Physics and Chemistry*, 3rd impression 2001, enlarged and revised edn. Elsevier, Amsterdam
- Stauffer D, Aharony A (1994) *Introduction to Percolation Theory*. Taylor, London
- Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1998) *The Theory of Critical Phenomena*. Clarendon Press, Oxford
- Binder K, Heermann DW (1997) *Monte Carlo Simulation in Statistical Physics*. Springer, Berlin
- Zia RKP, Schmittmann B (2007) Probability currents as principal characteristics in the statistical mechanics of non-equilibrium steady states. *J Stat Mech* P07012
- de Oliveira MM, Dickman R (2005) How to simulate the quasistationary state. *Phys Rev E* 71:016129
- Spitzer F (1970) Interaction of Markov processes. *Adv Math* 5:256–290
- Bouchaud JP, Comtet A, Georges A, Le Doussal P (1990) Classical Diffusion of a Particle in a One-Dimensional Random Force Field. *Ann Phys* 201:285–341
- Grassberger P (2002) Go with the winners: a general Monte Carlo strategy. *Comp Phys Comm* 147:64–70
- Pradhan P, Dhar D (2006) Sampling rare fluctuations of height in the Oslo ricepile model. *J Phys A* 40:2639–2650
- Bouchaud JP, Potters M (2003) *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, Cambridge
- De Menech M, Stella AL, Tebaldi C (1998) Rare events and breakdown of simple scaling in the Abelian sandpile model. *Phys Rev E* 58:R2677–R2680
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in C*, 2nd edn. Cambridge University Press, New York
- Knuth DE (1997) *The Art of Computer Programming*, vol 3. Seminumerical Algorithms, 2nd edn. Addison Wesley, Massachusetts
- Matsumoto M, Nishimura T (1998) Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudorandom Number Generator. *ACM Trans Model Comp Sim* 8:3–30
- Gentle JE (1998) *Random Number Generation and Monte Carlo Methods*. Springer, Berlin
- Ferrenberg AM, Landau DP (1992) Monte Carlo Simulations: Hidden Errors from “Good” Random Number Generators. *Phys Rev Lett* 69:3382–3384
- Efron B (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia
- Berg BA (1992) Double Jackknife bias-corrected estimators. *Comp Phys Com* 69:7–14
- Fisher ME (1967) The Theory of equilibrium critical phenomena. *Rep Prog Phys* 30:615–730
- Privman V, Hohenberg PC, Aharony A (1991) Universal Critical-Point Amplitude Relations. In: Domb C, Lebowitz JL (eds) *Phase Transitions and Critical Phenomena*, vol 14. Academic Press, New York, pp 1–134
- Christensen K, Corral A, Frette V, Feder J, Jøssang T (1996) Tracer Dispersion in a Self-Organized Critical System. *Phys Rev Lett* 77:107–110
- Pruessner G, Jensen HJ (2002) Broken scaling in the forest-fire model. *Phys Rev E* 65:056707–1–8. Preprint cond-mat/0201306
- Wegner FJ (1972) Corrections to Scaling Laws. *Phys Rev B* 5:4529–4536
- Dowd K, Severance C (1998) *High Performance Computing*, 2nd edn. O’Reilly, Sebastopol
- Pfeuty P, Toulouse G (1977) *Introduction to the Renormalization Group and to Critical Phenomena*. Wiley, Chichester
- Ferrenberg AM, Landau DP, Binder K (1991) Statistical and Systematic Errors in Monte Carlo Sampling. *J Stat Phys* 63:867–882
- Milchev A, Binder K, Heermann DW (1986) Fluctuations and Lack of Self-Averaging in the Kinetics of Domain Growth. *Z Phys B* 63:521–535

34. Christensen K, Moloney NR (2005) Complexity and Criticality. Imperial College Press, London
35. Pruessner G, Jensen HJ (2004) Efficient algorithm for the forest fire model. Phys Rev E 70:066707–1–25. Preprint cond-mat/0309173

### Books and Reviews

- Anderson TW (1964) The Statistical Analysis of Time Series. Wiley, London
- Barenblatt GI (1996) Scaling, self-similarity, and intermediate asymptotics. Cambridge University Press, Cambridge
- Berg BA (2004) Markov Chain Monte Carlo Simulations and Their Statistical Analysis. World Scientific, Singapore
- Brandt S (1998) Data Analysis. Springer, Berlin
- Jensen HJ (1998) Self-Organized Criticality. Cambridge University Press, New York
- Liggett TM (2005) Stochastic Interacting Systems: Contact, Voter and Exclusion Processes. Springer, Berlin
- Marro J, Dickman R (1999) Nonequilibrium Phase Transitions in Lattice Models. Cambridge University Press, New York
- Newman MEJ, Barkema GT (1999) Monte Carlo Methods in Statistical Physics. Oxford University Press, New York
- Stanley HE (1971) Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, New York

## Probability Distributions in Complex Systems

DIDIER SORNETTE

Department of Management, Technology and Economics, ETH Zurich, Zurich, Switzerland

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 The Fascination with Power Laws  
 Testing for Power Law Distributions in your Data  
 Beyond Power Laws: “Kings”  
 Future Directions  
 Bibliography

### Glossary

**Complex system** A system with a large number of mutually interacting parts, often open to its environment, which self-organizes its internal structure and its dynamics with novel and sometimes surprising macroscopic “emergent” properties.

**Criticality (in physics)** A state in which spontaneous fluctuations of the order parameter occur at all scales, leading to diverging correlation length and susceptibility of the system to external influences.

**Power law distribution** A specific family of statistical distribution appearing as a straight line in a log-log plot; exhibits the property of scale invariance and therefore does not possess characteristic scales.

**Self-organized criticality** Occurs when the system dynamics are attracted spontaneously, without any obvious need for parameter tuning to a critical state with infinite correlation length and power law statistics.

**Stretched-exponential distribution** A specific family of sub-exponential distribution interpolating smoothly between the exponential distribution and the power law family.

### Definition of the Subject

This core article for the Encyclopedia of Complexity and System Science (Springer Science) reviews briefly the concepts underlying complex systems and probability distributions. The latter are often taken as the first quantitative characteristics of complex systems, allowing one to detect the possible occurrence of regularities providing a step toward defining a classification of the different levels of organization (the “universality classes”). A rapid survey covers the Gaussian law, the power law and the stretched exponential distributions. The fascination for power laws is then explained, starting from the statistical physics approach to critical phenomena, out-of-equilibrium phase transitions, self-organized criticality, and ending with a large, but not exhaustive, list of mechanisms leading to power law distributions. A checklist for testing and qualifying a power law distribution from data is described in seven steps. This essay enlarges the description of distributions by proposing that “kings”, i.e., events even beyond the extrapolation of the power law tail, may reveal information which is complementary and perhaps sometimes even more important than the power law distribution. We conclude with a list of future directions.

### Introduction

#### Complex Systems

The study of out-of-equilibrium dynamics (e.g. dynamical phase transitions) and of heterogeneous systems (e.g., spin-glasses) has progressively made popular in physics the concept of complex systems and the importance of systemic approaches: systems with a large number of mutually interacting parts, often open to their environment, self-organize their internal structure and their dynamics with novel and sometimes surprising macroscopic (“emergent”) properties. The complex system approach,

which involves “seeing” inter-connections and relationships, i. e., the whole picture as well as the component parts, has become pervasive in modern control of engineering devices and business management. It also plays an increasing role in most of the scientific disciplines, including biology (biological networks, ecology, evolution, origin of life, immunology, neurobiology, molecular biology, etc.), geology (plate-tectonics, earthquakes and volcanoes, erosion and landscapes, climate and weather, environment, etc.), economics and social sciences (cognition, distributed learning, interacting agents, etc.). There is a growing recognition that progress in most of these disciplines, in many of the pressing issues for our future welfare as well as for the management of our everyday life, will need such a systemic complex system and multidisciplinary approach.

A central property of a complex system is the possible occurrence of coherent large-scale collective behaviors with very rich structure, resulting from the repeated nonlinear interactions among its constituents: the whole turns out to be much more than the sum of its parts. Most complex systems around us exhibit rare and sudden transitions that occur over time intervals which are short compared to the characteristic time scales of their prior evolution. Such extreme events express more than anything else the underlying “forces” usually hidden by almost perfect balance and thus provide the potential for a better scientific understanding of complex systems. These crises have fundamental societal impacts and range from large natural catastrophes such as earthquakes, volcanic eruptions, hurricanes and tornadoes, landslides, avalanches, lightning strikes, and catastrophic events of environmental degradation, to the failure of engineering structures, crashes in the stock market, social unrest leading to large-scale strikes and upheaval, economic drawdowns on national and global scales, regional power blackouts, traffic gridlock, diseases and epidemics, etc.

Given the complex dynamics of these systems, a first standard attempt to quantify and classify the characteristics and the possible different regimes consists of

1. identifying discrete events,
2. measuring their sizes,
3. constructing their probability distribution.

The interest in probability distributions in complex systems has the following roots.

- They offer a natural metric of the relative rate of occurrence of small versus large events, and thus of the associated risks.
- As such, they constitute essential components of risk assessment and prerequisites of risk management.
- Their mathematical form can provide constraints and guidelines to identify the underlying mechanisms at their origin and thus at the origin of the behavior of the complex system under study.
- This improved understanding may lead to better forecasting skills, and even to the option (or illusion (?)) of (a certain degree of) control [1,2].

## Probability Distributions

Let us first establish some notation and vocabulary. Consider a process  $X$  whose outcome is a real number. The probability density function  $P(x)$  of  $X$  (also called probability distribution or pdf) is such that the probability that  $X$  is found in a small interval  $\Delta x$  around  $x$  is  $P(x)\Delta x$ . The probability that  $X$  is between  $a$  and  $b$  is therefore given by the integral of  $P(x)$  between  $a$  and  $b$ :

$$P(a < X < b) = \int_a^b P(x)dx. \quad (1)$$

The pdf  $P(x)$  depends on the units used to quantify the variable  $x$  and has the dimension of the inverse of  $x$ , such that  $P(x)\Delta x$ , being a probability, i. e., a number between 0 and 1, is dimensionless. In a change of variable, say  $x \rightarrow y = f(x)$ , the probability is invariant. Thus, the invariant quantity is the probability  $P(x)\Delta x$  and not the pdf  $P(x)$ . We thus have

$$P(x)\Delta x = P(y)\Delta y, \quad (2)$$

leading to  $P(y) = P(x)|df/dx|^{-1}$ , taking the limit of infinitesimal intervals. By definition,  $P(x) \geq 0$ . It is normalized,  $\int_{x_{\min}}^{x_{\max}} P(x)dx = 1$ , where  $x_{\min}$  and  $x_{\max}$  (often  $\pm\infty$ ) are the smallest and largest possible values for  $x$ , respectively.

The empirical estimation of the pdf  $P(x)$  is usually plotted with the horizontal axis scaled as a graded series for the measure under consideration (the magnitude of the earthquakes, etc.) and the vertical axis scaled for the number of outcomes or measures in each interval of horizontal value (the number of earthquakes of magnitude between 1 and 2, between 2 and 3, etc.). This implies a “binning” into small intervals. If the data is sparse, the number of events in each bin becomes small and can fluctuate, leading to a poor representation of the data. In this case, it is useful to construct the cumulative distribution  $P_{\leq}(x)$  defined by

$$P_{\leq}(x) = P(X \leq x) = \int_{-\infty}^x P(y)dy, \quad (3)$$

which is much less sensitive to fluctuations.  $P_{\leq}(x)$  gives the fraction of events with values less than or equal to  $x$ .  $P_{\leq}(x)$  increases monotonically with  $x$  from 0 to 1. Similarly, we can define the so-called complementary cumulative (or survivor) distribution  $P_{>}(x) = 1 - P_{\leq}(x)$ .

For random variables which take only discrete values  $x_1, x_2, \dots, x_n$ , the pdf is made of a discrete sum of Dirac functions  $(1/n)[\delta(x - x_1) + \delta(x - x_2) + \dots + \delta(x - x_n)]$ . The corresponding cumulative distribution function (cdf)  $P_{\leq}(x)$  is a staircase. There are also more complex distributions made of a continuous cdf but which are singular with respect to the Lebesgue measure  $dx$ . An example is the Cantor distribution constructed from the Cantor set (see for instance Chap. 5 in [3]). Such a singular cdf is continuous but has its derivative which is zero almost everywhere: the pdf does not exist (see e. g. [4]).

### Brief Survey of Probability Distributions

Statistical physics is rich with probability distributions. The most famous is the Boltzmann distribution, which describes the probability that the configuration of a system in thermal equilibrium has a given energy. Its extension to out-of-equilibrium systems is the subject of intense scrutiny [5]; see also Chap. 7 in [3] and references therein. Special cases include the Maxwell–Boltzmann distribution, the Bose–Einstein distribution and the Fermi–Dirac distribution.

In the quest to characterize complex systems, two distributions have played a leading role: the normal (or Gaussian) distribution and the power law distribution. The Gaussian distribution is the paradigm of the “mild” family of distributions. In contrast, the power law distribution is the representative of the “wild” family. The contrast between “mild” and “wild” is illustrated by the following questions:

- What is the probability that someone has twice your height? Essentially zero! The height, weight and many other variables are distributed with “mild” pdfs with a well-defined typical value and relatively small variations around it. The Gaussian law is the archetype of “mild” distributions.
- What is the probability that someone has twice your wealth? The answer of course depends somewhat on your wealth but in general, there is a non-vanishing fraction of the population twice, ten times or even one hundred times as wealthy as you are. This was noticed at the end of the 19th century by Pareto, after whom the Pareto law has been named, which describes the power law distribution of wealth [6,7], a typical example of “wild” distributions.

**The Normal (or Gaussian) Distribution** The expression of the Gaussian probability density function of a random variable  $x$  with mean  $x_0$  and standard deviation  $\sigma$  reads

$$P_G(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right) \quad \text{defined for } -\infty < x < +\infty. \quad (4)$$

The importance of the normal distribution as a model of quantitative phenomena in the natural and social sciences can be in large part attributed to the central limit theorem. Many measurements of physical as well as social phenomena can be well approximated by the normal distribution. While the mechanisms underlying these phenomena are often unknown, the use of the normal model can be theoretically justified by assuming that many small, independent effects additively contribute to each observation. The Gaussian distribution is also justified as the most parsimonious choice in absence of information other than just the mean and the variance: it maximizes the information entropy among all distributions with known mean and variance. As a result of the central limit theorem, the normal distribution is the most widely used family of distributions in statistics and many statistical tests are based on the assumption of asymptotic normality of the data. In probability theory, the standard Gaussian distribution arises as the limiting distribution of a large class of distributions of random variables (with suitable centering and normalization) characterized by a finite variance, which is nothing but the statement of the central limit theorem (see Chapter 2 in [3]).

At the beginning of the twenty-first century, when power laws are often taken as the hallmark of complexity, it is interesting to reflect on the fact that the previous giants of science in the eighteenth and nineteenth centuries (Halley, Laplace, Quetelet, Maxwell and so on) considered that the Gaussian distribution expressed a kind of universal law of nature and of society. In particular, the Belgian astronomer Adolphe Quetelet was instrumental in popularizing the statistical regularities discovered by Laplace in the frame of the Gaussian distribution, which influenced the likes of John Herschel and John Stuart Mill and led Comte to define the concept of “social physics.”

**The Power Law Distribution** A probability distribution function  $P(x)$  exhibiting a power law tail is such that

$$P(x) \propto \frac{C_\mu}{x^{1+\mu}}, \quad \text{for } x \text{ large}, \quad (5)$$

possibly up to some large limiting cut-off. The exponent  $\mu$  (also referred to as the “index”) characterizes the nature of



the tail: for  $\mu < 2$ , one speaks of a “heavy tail” for which the variance is theoretically not defined. For power laws, the scale factor  $C_\mu$  plays a role analogous to the role of variance in Gaussian distributions (see Chapter 4 in [3]). In particular, it enjoys the additivity property: the scale factor of the distribution of the sum of several independent random variables, each with a distribution exhibiting a power law tail with the same exponent  $\mu$ , is equal to the sum of the scale factors characterizing each distribution of each random variable in the sum.

A more general form is

$$P(x) \propto \frac{L(x)}{x^{1+\mu}}, \quad \text{for } x \text{ large}, \quad (6)$$

where  $L(x)$  is a slowly varying function defined by  $\lim_{x \rightarrow \infty} L(tx)/L(x) = 1$  for any finite  $t$  (typically,  $L(x)$  is a logarithm  $\ln(x)$  or power of a logarithm such as  $(\ln(x))^n$  with  $n$  finite). In mathematical language, a function such as (6) is said to be “regularly varying”. This more general form means that the power law regime is an asymptotic statement holding only as a better and better approximation as one considers larger and larger  $x$  values.

Power laws obey the symmetry of scale invariance, that is, they verify the following defining property that, for an arbitrary real number  $\lambda$ , there exists a real number  $\gamma$  such that

$$P(x) = \gamma P(\lambda x), \quad \forall x. \quad (7)$$

Obviously,  $\gamma = \lambda^{1+\mu}$ . The relation (7) means that the ratio of the probabilities of occurrence of two sizes  $x_1$  and  $x_2$  depends only on their ratio  $x_1/x_2$  and not on their absolute values. For instance, according to the Zipf law ( $\mu = 1$ ) for the distribution of city sizes, the ratio of the number of cities with more than 1 million inhabitants to those with more than 100,000 persons is the same as the ratio of the number of cities with more than 100,000 inhabitants to those with more than 10,000 persons, both ratios being equal to 1/10. The symmetry of scale invariance (7) extends to the space of functions the concept of scale invariance which characterizes fractal geometric objects.

It should be stressed that, when they exhibit a power-law-like shape, most empirical distributions do so only over a finite range of event sizes, either bounded between a lower and an upper cut-off [8,9,10,11], or above a lower threshold, i.e., only in the tail of the observed distribution [12,13,14,15]. Power law distributions and, more generally, regularly varying distributions remain robust functional forms under a large number of operations, such as linear combinations, products, minima, maxima, order statistics, and powers, which may also explain their ubiquity and attractiveness. Jessen and Mikosch [16] give the

conditions under which transformations of power law distributions are also regularly varying, possibly with a different exponent (see also Sect. 4.4 in [3] for an heuristic presentation of similar results).

**The Stretched Exponential Distribution** The so-called stretched exponential (SE) distribution has been found to be a versatile intermediate distribution interpolating between “thin tail” (Gaussian, exponential, ...) and very “fat tail” distributions. In particular, Laherrère and Sornette [17] have found that several examples of fat-tailed distributions in the natural and social sciences, often considered to be good examples of power laws, could sometimes be represented as well as or even better by an SE distribution. Malevergne et al. [18] present systematic statistical tests comparing the SE family with the power law distribution in the context of financial return distributions. The SE family is defined by the following expression for the survival distribution (also called the complementary cumulative distribution function):

$$\mathcal{P}_{\geq u}(x) = 1 - \exp \left[ - \left( \frac{x}{d} \right)^c + \left( \frac{u}{d} \right)^c \right], \quad \text{for } x \geq u. \quad (8)$$

The constant  $u$  is a lower threshold that can be changed to increasingly emphasize the tail of the distribution as  $u$  is increased. The structural exponent  $c$  controls the “thin” versus “heavy” nature of the tail.

1. For  $c = 2$ , the SE distribution (8) has the same asymptotic tail as the Gaussian distribution.
2. For  $c = 1$ , expression (8) recovers the pure exponential distribution.
3. For  $c < 1$ , the tail of  $\mathcal{P}_u(x)$  is fatter than an exponential, and corresponds to the regime of sub-exponentials (see Chapter 6 in [3]).
4. For  $c \rightarrow 0$  with

$$c \cdot \left( \frac{u}{d} \right)^c \rightarrow \mu, \quad (9)$$

the SE distribution converges to the Pareto distribution with tail exponent  $\mu$ .

Indeed, we can write

$$\begin{aligned} & \frac{c}{d^c} \cdot x^{c-1} \cdot \exp \left( - \frac{x^c - u^c}{d^c} \right) \\ &= c \left( \frac{u}{d} \right)^c \cdot \frac{x^{c-1}}{u^c} \exp \left[ - \left( \frac{u}{d} \right)^c \cdot \left( \left( \frac{x}{u} \right)^c - 1 \right) \right], \\ &\simeq \mu \cdot x^{-1} \exp \left[ -c \left( \frac{u}{d} \right)^c \cdot \ln \frac{x}{u} \right], \quad \text{as } c \rightarrow 0 \\ &\simeq \mu \cdot x^{-1} \exp \left[ -\mu \cdot \ln \frac{x}{u} \right], \\ &\simeq \mu \frac{u^\mu}{x^{\mu+1}}, \end{aligned} \quad (10)$$

which is the pdf of the Pareto power law model with tail index  $\mu$ . This implies that, as  $c \rightarrow 0$ , the characteristic scale  $d$  of the SE model must also go to zero with  $c$  to ensure its convergence towards the Pareto distribution.

This shows that the Pareto model can be approximated with any desired accuracy on an arbitrary interval ( $u > 0, U$ ) by the (SE) model with parameters ( $c, d$ ) satisfying Eq. (9) where the arrow is replaced by an equality. The limit  $c \rightarrow 0$  provides any desired approximation to the Pareto distribution uniformly on any finite interval ( $u, U$ ). This deep relationship between the SE and power law models allows us to understand why it can be very difficult to decide, on a statistical basis, which of these models best fits the data [17,18]. This insight can be made rigorous to develop a formal statistical test of the (SE) hypothesis versus the Pareto hypothesis [18,19].

From a theoretical viewpoint, this class of distributions (8) is motivated in part by the fact that large deviations of multiplicative processes are generically distributed with stretched exponential distributions [20]. Stretched exponential distributions are also parsimonious examples of the important subset of sub-exponentials, that is, of the general class of distributions decaying slower than an exponential [21]. This class of sub-exponentials share several important properties of heavy-tailed distributions [22] not shared by exponentials or distributions decreasing faster than exponentials: for instance, they have “fat tails” in the sense of the asymptotic probability weight of the maximum compared with the sum of large samples [4] (see also Chaps 1 and 6 in [3]).

Notwithstanding their fat-tailness, stretched exponential distributions have only finite moments, in contrast with regularly varying distributions for which moments of order equal to or larger than the tail index  $\mu$  are not defined. However, they do not admit an exponential moment, which leads to problems in the reconstruction of the distribution from the knowledge of their moments [23]. In addition, the existence of all moments is an important property allowing for an efficient estimation of any high-order moment, since it ensures that the estimators are asymptotically Gaussian. In particular, for stretched-exponentially distributed random variables, the variance, skewness and kurtosis can be accurately estimated, contrarily to random variables with regularly varying distribution with tail index smaller than about 5.

### The Fascination with Power Laws

Probability distribution functions with a power law dependence in terms of event or object sizes seem to be ubiquitous statistical features of natural and social systems. It

has repeatedly been argued that such an observation relies on an underlying self-organizing mechanism, and therefore power laws should be considered as the statistical imprints of complex systems. It is often claimed that the observation of a power law relation in data often points to specific kinds of mechanisms at its origin, that can often suggest a deep connection with other, seemingly unrelated systems. In complex systems, the appearance of power law distributions is often thought to be the signature of hierarchy and robustness. In the last two decades, such claims have been made, for instance, for earthquakes, weather and climate changes, solar flares, the fossil record, and many other systems, to promote the relevance of self-organized criticality as an underlying mechanism for the organization of complex systems [24]. This claim is often unwarranted as there are many non-self-organizing mechanisms that produce power law distributions [3,25,26,27].

Research on the origins of power law relations and efforts to observe and validate them in the real world are extremely active in many fields of modern science, including physics, geophysics, biology, medical sciences, computer science, linguistics, sociology and economics. This section briefly summarizes the present understanding.

### Statistical Physics in General and the Theory of Critical Phenomena

The study of critical phenomena in statistical physics suggests that power laws emerge close to special critical or bifurcation points separating two different phases or regimes of the system. In systems at thermodynamic equilibrium modeled by general spin models, renormalization group theory [28] has demonstrated the existence of universality, so that diverse systems exhibit the same critical exponents and identical scaling behavior as they approach criticality, i. e., they share the same fundamental macroscopic properties. For instance, the behavior of water and CO<sub>2</sub> at their boiling points at a certain critical pressure and that of a magnet at its Curie point fall in the same universality class because they can be characterized by order parameter  $s$  with the same symmetries and dimensions in the same space dimension. In fact, almost all material phase transitions are described by a small set of universality classes.

From this perspective, the fascination with power laws reflects the fact that they characterize the many coexisting and delicately interacting scales at a critical point. The existence of many scales leading to complex geometrical properties is often associated with fractals [12]. While it is true that critical points and fractals share power law relations, power law relations and power law distributions

are not the same. The latter, which is the subject of this essay, describes the probability density function or frequency of occurrence of objects or events, such as the frequency of earthquakes of a given magnitude range. In contrast, power law relations between two variables (such as the magnetization and temperature in the case of the Curie point of a magnet) describe a functional abstraction belonging to or characteristic of these two variables. Both power law relations and power law distributions can result from the existence of a critical point. A simple example in percolation is (i) the power law dependence of the size of the larger cluster as a function of the distance from the percolation threshold and (ii) the power law distribution of cluster sizes at the percolation threshold [29].

### Out-of-Equilibrium Phase Transition and Self-Organized Critical Systems (SOC)

In the broadest sense, self-organized criticality (SOC) refers to the spontaneous organization of a system driven from the outside into a globally stationary state, which is characterized by self-similar distributions of event sizes and fractal geometrical properties. This stationary state is dynamical in nature and is characterized by statistical fluctuations, which are generically referred to as “avalanches.”

The term “self-organized criticality” contains two parts. The word “criticality” refers to the state of a system at a critical point at which the correlation length and the susceptibility become infinite in the infinite size limit as in the preceding section. The label “self-organized” is often applied indiscriminately to pattern formation among many interacting elements. The concept is that the structuration, the patterns and large scale organization appear spontaneously. The notion of self-organization refers to the absence of control parameters.

In this class of mechanisms, where the critical point is the attractor, the situation becomes more complicated as the number of universality classes proliferates. In particular, it is not generally well-understood why sometimes local details of the dynamics may change the macroscopic properties completely while in other cases, the universality class is robust. In fact, the more we learn about complex out-of-equilibrium systems, the more we realize that the concept of universality developed for critical phenomena at equilibrium has to be enlarged to embody a more qualitative meaning: the critical exponents defining the universality classes are often very sensitive to many (but not all) details of the models [30].

Of course, one of the hailed hallmarks of SOC is the existence of power law distributions of “avalanches” and of other quantities [3,24,31].

### Non-exhaustive List of Mechanisms Leading to Power Law Distributions

There are many physical and mathematical mechanisms that generate power law distributions and self-similar behavior. Understanding how a mechanism is affected by microscopic laws constitutes an active field of research. We can propose the following non-exhaustive list of mechanisms that have been found to operate in different complex systems, and which can lead to power law distribution of avalanches or cluster sizes. For most of these mechanisms, we refer the reader to Chaps. 14 and 15 in [3] and to [27,32] for detailed explanations and the relevant bibliography. However, some of the mechanisms mentioned here have not been reviewed in these three references and are thus new to the list developed in particular in [3]. We should also stress that some of the mechanisms in this list are actually different incarnations of the same underlying idea (for instance preferential attachment which is a re-discovery of the Yule process, see [33] for an informative historical account).

1. percolation, fragmentation and other related processes,
2. directed percolation and its universality class of so-called “contact processes,”
3. cracking noise and avalanches resulting from the competition between frozen disorder and local interactions, as exemplified in the random field Ising model, where avalanches result from hysteretic loops [34],
4. random walks and their properties associated with their first passage statistics [35] in homogeneous as well as in random landscapes,
5. flashing annihilation in Verhulst kinetics [36],
6. sweeping of a control parameter towards an instability [25,37],
7. proportional growth by multiplicative noise with constraints (the Kesten process [38] and its generalization, for instance in terms of generalized Lotka–Volterra processes [39]), whose ancestry can be traced to Simon and Yule,
8. competition between multiplicative noise and birth-death processes [40],
9. growth by preferential attachment [32],
10. exponential deterministic growth with random times of observations (which gives the Zipf law) [41],
11. constrained optimization with power law constraints (HOT for “highly optimized tolerant”),
12. control algorithms, which employ optimal parameter estimation based on past observations, shown to generate broad power law distributions of fluctuations

and of their corresponding corrections in the control process [42,43],

13. on-off intermittency as a mechanism for power law pdf of laminar phases [44,45],
14. self-organized criticality which comes in many flavors as explained in Chapter 15 of [3]:
  - cellular automata sandpiles with and without conservation laws,
  - systems made of coupled elements with threshold dynamics,
  - critical de-synchronization of coupled oscillators of relaxation,
  - nonlinear feedback of the order parameter onto the control parameter
  - generic scale invariance,
  - mapping onto a critical point,
  - extremal dynamics.

If there is one lesson to extract from this impressive list it is that, when observing an approximate linear trend in the log-log plot of some data distribution, one should refrain from jumping to hasty conclusions on the implications of this approximate power law behavior. Another lesson is that power laws appear to be so ubiquitous perhaps because many roads lead to them!

### Testing for Power Law Distributions in your Data

Although power law distributions are attractive for their simplicity (they are straight lines on log-log plots) and may be justified from theoretical reasons as discussed above, demonstrating that data do indeed follow a power law distribution requires more than simple fitting. Indeed, several alternative functional forms can appear to follow a power law form over some extent, such as stretched exponentials and log-normal distributions. Thus, validating that a given distribution is a power law is not easy and there is no silver bullet.

Clauset et al. [46] have recently summarized some statistical techniques for making accurate parameter estimates for power-law distributions based on maximum likelihood methods and the Kolmogorov–Smirnov statistic. They illustrate these statistical methods on 24 real-world data sets from a range of different disciplines. In some cases, they find that power laws are consistent with the data while in others the power law is ruled out. The log-likelihood ratio that they propose is however not warranted for non-nested models [47]

Here, we offer some advice for the characterization of a power law distribution as a possible adequate representation of a given data set. We emphasize good sense and practical aspects.

1. **Survivor distribution** First, the survival distribution should be constructed using raw data by ranking the values in increasing order. Then, rank versus values gives immediately a non-normalized survival distribution. The advantage of this construction is that it does not require binning or kernel estimation, which is a delicate art, as we have alluded to.
  2. **Probability density function** The previous construction of the complementary cumulative (or survivor) distribution function should be complemented with that of the density function. Indeed, it is well-known that the cumulative distribution, being a “cumulative” integral of the density function as its name indicates, may be contaminated by disturbances at one end of the density function, leading to rather long cross-overs that may hide or perturb the power law. For instance, if the generating density distribution is a power law truncated by an exponential, as found for critical systems not exactly at their critical point or in the presence of finite-size effects [48], the power law part of the cumulative distribution will be strongly distorted leading to a spurious estimation of the exponent  $\mu$ . This problem can be in large part alleviated by constructing the pdf using binning or, even better, kernel methods (see the very readable article [49] and references therein). By testing and comparing the survival and the probability density distributions, one obtains either a confirmation of power law scaling or an understanding of the origin(s) of the deviations from the power law.
  3. **Structural analysis by visual inspection** Given that these first two steps have been performed, we recommend a preliminary visual exploration by plotting the survival and density distributions in (i) linear-linear coordinates, (ii) log-linear coordinates (linear abscissa and logarithmic ordinate) and (iii) log-log coordinates (logarithmic abscissa and logarithmic ordinate). The visual comparison between these three plots provides a fast and intuitive view of the nature of the data.
    - A power law distribution will appear as a convex curve in the linear-linear and log-linear plots and as a straight line in the log-log plot.
    - A Gaussian distribution will appear as a bell-shaped curve in the linear-linear plot, as an inverted parabola in the log-linear plot and as a strongly concave sharply falling curve in the log-log plot.
    - An exponential distribution will appear as a convex curve in the linear-linear plot, as a straight line in the log-linear plot and as a concave curve in the log-log plot.
- Having in mind the shape of these three reference distributions in these three representations provides fast

and useful reference points to classify the unknown distribution under study. For instance, if the log-linear plot shows a convex shape (upward curvature), we can conclude that the distribution has a tail fatter than an exponential. Then, the log-log plot will confirm if a power law is a reasonable description. If the log-log plot shows a downward curvature (concave shape), together with the information that the log-linear plot shows a convex shape, we can conclude that the distribution has a tail fatter than an exponential but thinner than a power law. For example, it could be a gamma distribution ( $\sim x^n \exp[-x/x_0]$  with  $n > 0$ ) or a stretched distribution (expression (8) with  $c < 1$ ). Only more detailed quantitative analysis will allow one to refine the diagnostic, often with less-than-definite conclusions (see as an illustration the detailed statistical analysis comparing the power law to the stretched exponential distributions to describe the distribution of financial returns [18]).

The deviations from linearity in the log-log plot suggest the boundaries within which the power law regime holds. We say “suggest,” as a visual inspection is only a first step which can be actually misleading. While we recommend a first visual inspection, it is only a first indication, not a proof. It is a necessary step to convince oneself (and the reviewers and journal editors) but certainly not a sufficient condition. It is a standard rule of thumb that power law scaling is thought to be meaningful if it holds over at least two to three decades on both axes and is bracketed by deviations on both sides whose origins can be understood (for instance, due to insufficient sampling and/or finite-size effects).

As an illustration of the potential errors stemming from visual inspection, we refer to the discussion of Sornette et al. [50], on the claim of Pacheco et al. [51] of the existence of a break in the Gutenberg–Richter distribution of earthquake magnitudes at  $m = 6.4$  for California. This break was claimed to reveal the finiteness of the crust thickness according to Pacheco et al. [51]. This claim has subsequently been shown to be unsubstantiated, as the Gutenberg–Richter law (which is a power law when expressed in earthquake energies or seismic moments) seems to remain valid up to magnitudes of 7.5 in California and up to magnitude about 8–8.5 worldwide. This visual break at  $m = 6.4$  turned out to be just a statistical deviation, completely expected from the nature of power law fluctuations [15,52].

4. **OLS fitting** The next step is often to perform an ordinary least-square (OLS) regression of the data (survival distribution or kernel-reconstructed density) on the logarithm of the variables, in order to estimate the

parameters of the power law. These parameters are the exponent  $\mu$ , the scale factor  $C_\mu$  and possibly an upper threshold or other parameters controlling the cross-over to other behaviors outside the scaling regime. Using logarithms ensures that all the terms in the sum of squares over the different data points contribute approximately similarly in the OLS. Otherwise, without logarithms, given the large range of values spanned by a typical power law distribution, a relative error of say 1% around a value of the order of  $10^4$  would have a weight in the sum ten thousand times larger than the weight due to the same relative error of 1% around a value of the order of  $10^2$ , biasing the estimation of the parameters towards fitting preferentially the large values. In addition, in logarithm units, the estimation of the exponent  $\mu$  of a power law constitutes a linear problem which is solved analytically. When performing the OLS estimation on the survival distribution, it is optimal to shift the ranks by  $1/2$  [53]. With this improvement, the OLS method is typically more robust to deviations from a pure power law form than the Hill estimator discussed below.

5. **Maximum likelihood estimation** Using an OLS method to estimate the parameters of a power law assumes implicitly that the distribution of the deviations from the power law (actually the difference between the logarithm of the data and the logarithm of the power law distribution) are normally distributed. This may not be a suitable approximation. An estimation which removes this assumption consists in using the likelihood method, in which the parameters of the power law are chosen so as to maximize the likelihood function. When the data points are independent, the likelihood function is nothing but the product  $\prod_{i=1}^N P(x_i)$  over the  $N$  data points  $x_1, x_2, \dots, x_N$  of the power law distribution  $P(x)$ . In this case, the exponent  $\mu$  which maximizes this likelihood (or equivalently and more conveniently its logarithm called the log-likelihood) is called the Hill estimator [54]. It reads

$$\frac{1}{\mu} = \frac{1}{n} \sum \ln \left[ \frac{x_j}{x_{\min}} \right], \quad (11)$$

where  $x_{\min}$  is the smallest value among the  $n$  values used in the data set for the estimation of  $\mu$ . Since power laws are often asymptotic characteristics of the tail, it is appropriate not to use the full data set but only the upper tail with data values above a lower threshold. Then, plotting  $1/\mu$  or  $\mu$  as a function of the lower threshold usually provides a good sense of the existence of a power law regime: one should expect an approximate stability of  $1/\mu$  over some scaling regime. Note



that the Hill estimator provides an unbiased estimate of  $1/\mu$  while  $\mu$  obtained by inverting  $1/\mu$  is slightly biased (see e. g., Chapter 6 in [3]). We refer to [55,56] for improved versions and procedures of the Hill estimator which deal with finite ranges and dependence.

6. **Non-parametric methods** Methods testing for a power law behavior in a given empirical distribution which are not parametric and not sensitive to the value of the exponent provide useful complements of the above fitting and parametric estimation approaches. Pisarenko et al. [57] and Pisarenko and Sornette [58] have developed new statistics such that a power law behavior is associated with a zero value of the statistics *independently of the numerical value of the exponent  $\mu$*  and with a non-zero value otherwise. Plotting these statistics as a function of the lower threshold of the data sample allows one to detect subtle deviations from a pure power law. Lasocki [59] and Lasocki and Papadimitriou [60] have developed another non-parametric approach to detect deviations from a power law, the smoothed bootstrap test for multimodality, which makes it possible to test the complexity of the distribution without specifying any particular probabilistic model. The method relies on testing the hypotheses that the number of modes or the number of bumps exhibited by the distribution function equal 1. Rejection of one of these hypotheses indicate that the distribution has more complexity than described by a simple power law.

Once the evidence for a power law distribution has been reasonably demonstrated, the most difficult task remains: finding a mechanism and model which can explain the data. Note that the term “explain” refers to different meanings depending on the expert you are speaking to. For a statistician, having been unable to reject the power law function (5) given the data amounts to saying that the power law model “explains” the data. The emphasis of the statistician will be on refining parametric and non-parametric procedures to test the way the power law “fits” or deviates from the empirical data. In contrast, a physicist or a natural scientist sees this as only a first step, and attributes the word “explain” to the stage where a mechanism described in terms of a more fundamental process or first principles can derive the power law. But even among natural scientists, there is no consensus on what is a suitable “explanation.” The reason stems from the different cultures and levels of study in different fields, well addressed in the famous paper “More is different” by Anderson [61]: a suitable explanation for a physicist will frustrate a chemist whose explanation, in turn, will not satisfy a biologist. Each scientific discipline’s concepts are

anchored in its characteristic fundamental scientific level, which provides the underpinning for the next scientific level of description (think for instance of the hierarchy: physics  $\rightarrow$  chemistry  $\rightarrow$  molecular biology  $\rightarrow$  cell biology  $\rightarrow$  animal biology  $\rightarrow$  ethology  $\rightarrow$  sociology  $\rightarrow$  economics  $\rightarrow$  ...).

Once a model at a given scientific description level has been proposed, the action of the model on inputs gives outputs which are compared with the data. Verifying that the model, inspired by the preliminary power law evidence, adequately fits this power law is a first step. Unfortunately, much too often, scientists stop there and are happy to report that they have a model that fits their empirical power law data. This is not good science. Keeping in mind the many possible mechanisms at the origin of power law distributions reviewed above, a correct procedure is to run the candidate model to get other predictions that can themselves be put to the test. This validation is essential to determine the degree to which the model is an accurate representation of the real world from the perspective of its intended uses. Reviewing a large body of literature devoted to the problem of validation, Sornette et al. [62] have proposed a synthesis in which the validation of a given model is formulated as an iterative construction process that mimics the often implicit process occurring in the minds of scientists. Validation is nothing but the progressive build-up of trust in the model, based on testing the model against non-redundant novel experiments or data, that allows one to make a decision and act on that basis. The applications of the validation program to a cellular automaton model for earthquakes, to a multifractal random walk model for financial time series, to an anomalous diffusion model for solar radiation transport in the cloudy atmosphere, and to a computational fluid dynamics code for the Richtmyer–Meshkov instability, exemplify the importance of going beyond the simple qualification of a power law.

## Beyond Power Laws: “Kings”

### The Standard View

Power law distributions embody the notion that extreme events are not exceptional 9-sigma events (to refer to the terminology using the Gaussian bell curve and its standard deviation  $\sigma$  as the metric to quantify deviations from the mean). Instead, extreme events should be considered as rather frequent and part of the same organization as other events. In this view, a great earthquake is just an earthquake that started small... and did not stop; it is inherently unpredictable due to its sharing of all the properties and characteristics of smaller events (except for its size),

so that no genuinely informative precursor can be identified [63]. This is the view expounded by Bak and co-workers in their formulation of the concept of self-organized criticality [24,64]. In the following, we outline several promising directions of research that expand on these ideas.

### Self-Organized Criticality Versus Criticality

However, there are many suggestions that inherent unpredictability does not need to be the case. One argument is that criticality and self-organized criticality (SOC) can actually co-exist. The hallmark of criticality is the existence of specific precursory patterns (increasing susceptibility and correlation length) in space and time. Continuing with the example of earthquakes, the idea that a great earthquake could result from a critical phenomenon has been put forward by different groups, starting almost three decades ago [65,66,67]. Attempts to link earthquakes and critical phenomena find support in the evidence that rupture in heterogeneous media is similar to a critical phenomenon (see Chapter 13 of [3] and references therein). Also indicative is the often-reported observation of increased intermediate magnitude seismicity before large events [68,69]. An illustration of the coexistence of criticality and of SOC is found in a simple sandpile model of earthquakes on a hierarchical fault structure [70]. Here, the important ingredient is to take into account both non-linear dynamics and complex geometry. From the point of view of self-organized criticality, this is surprising news: large earthquakes do not lose their identity. In the model of Huang et al. [70], a large earthquake is different from a small one, a very different story than the one told by common SOC wisdom in which any precursory state of a large event is essentially identical to a precursory state of a small event and an earthquake does not “know” how large it will become. The difference comes from the absence of geometry in standard SOC models. Reintroducing geometry is essential. In models with hierarchical fault structures, one finds a degree of predictability of large events.

### Beyond Power Laws: Five Examples of “Kings”

Are power laws the whole story? The following examples suggest that some extreme events are even “wilder” than predicted by the extrapolation of power law distributions. They can be termed “outliers” or even better “kings” [17]. According to the definition of the Engineering Statistical Handbook [71], “An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.” Here, we follow Laherrère and Sornette [17] and use the term “king” to refer to events which

are even beyond the extrapolation of the fat tail distribution of the rest of the population.

- Material failure and rupture processes** There is now ample evidence that the distribution of damage events, for instance quantified by the acoustic emission radiated by micro-cracking in heterogeneous systems, is well-described by a Gutenberg–Richter like power law [72,73,74,75]. But consider now the energy released in the final global event rupturing the system in pieces! This release of energy is many, many times larger than the largest ever recorded event in the power law distribution. Material rupture exemplifies the co-existence of a power law distribution and a catastrophic event lying beyond the power law.
- Gutenberg–Richter law and characteristic earthquakes** In seismo-tectonics, the situation is muddy because of the difficulties with unambiguously defining the spatial domain of influence of a given fault. Researchers who have delineated a spatial domain surrounding a clearly mapped large fault claim to find a Gutenberg–Richter distribution up to a large magnitude region characterized by a bump or anomalous rate of large earthquakes. These large earthquakes have rupture lengths comparable with the fault length [76,77]. If proven valid, this concept of a characteristic earthquake provides another example in which a “king” co-exists with a power law distribution of smaller events. Others have countered that this bump disappears when removing the somewhat artificial partition of the data [78,79], so that the characteristic earthquake concept may be a statistical artifact. In this view, a particular fault may appear to have characteristic earthquakes, but the stress-shedding region, as a whole, behaves according to a pure scale-free power law distribution. Several theoretical models have been offered to support the idea that, in some seismic regimes, there is a co-existence between a power law and a large size regime (the “king” effect). Gil and Sornette [80] reported that this occurs when the characteristic rate for local stress relaxations is fast compared with the diffusion of stress within the system. The interplay between dynamical effects and heterogeneity has also been shown to change the Gutenberg–Richter behavior to a distribution of small events combined with characteristic system size events [81,82,83,84]. On the empirical side, progress should be made in testing the characteristic earthquake hypothesis by using the prediction of the models to identify independently of seismicity those seismic regions in which the king effect is expected. This remains to be done [85].

- Extreme king events in the pdf of turbulent velocity fluctuations** The evidence for kings does not require, and is not even synonymous in general with, the existence of a break or a bump in the distribution of event sizes. This point is well-illustrated in shell models of turbulence which are believed to capture the essential ingredient of these flows, while being amenable to analysis. Such “shell” models replace the three-dimensional spatial domain by a series of uniform onion-like spherical layers with radii increasing as a geometrical series  $1, 2, 4, 8, \dots, 2^n$  and communicating mostly with nearest neighbors. The quantity of interest is the distribution of velocity variations between two instants at the same position or between two points simultaneously. L’vov et al. [86] have shown that they could collapse the pdfs of velocity fluctuations for different scales only for small velocity fluctuations, while no scaling held for large velocity fluctuations. The conclusion is that the distributions of velocity increments seems to be composed of two regions, a region of so-called “normal scaling” and a domain of extreme events. They could also show that these extreme fluctuations of the fluid velocity correspond to intensive peaks propagating coherently (like solitons) over several shell layers with a characteristic bell-like shape, approximately independent of their amplitude and duration (up to a rescaling of their size and duration). One could summarize these findings by saying that “characteristic” velocity pulses decorate an otherwise scaling probability distribution function.
- Outliers and kings in the distribution of financial drawdowns** In a series of papers, Johansen and Sornette [87,88,89] have shown that the distribution of drawdowns in financial markets exhibits the coexistence of a fat tail with a characteristic regime with “kings” (called “outliers” in the papers). The analysis encompasses exchange markets (US dollar against the Deutsch Mark and against the Yen), the major world stock markets, the U.S. and Japanese bond markets and commodity markets. Here, drawdowns are defined as a continuous decrease in the value of the price at the close of each successive trading day. The results are found robust with using “coarse-grained drawdowns,” which allows for a certain degree of fuzziness in the definition of cumulative losses. Interestingly, the pdf of returns at a fixed time scale, usually the daily returns, does not exhibit any anomalous king behavior in the tail: the pdf of financial returns at fixed time scales seems to be adequately described by power law tails [90]. The interpretation proposed by Johansen and Sornette is that these drawdown kings are associated

with crashes, which occur due to a global instability of the market which amplifies the normal behavior via strong positive feedback mechanisms [91].

- Paris as the king in the Zipf distribution of French city sizes** Since Zipf [92], it is well-documented that the distribution of city sizes (measured by the number of inhabitants) is, in many countries, a power law with an exponent  $\mu$  close to 1. France is not an exception as it exhibits a nice power law distribution of city sizes... except for Paris which is completely out of range, a genuine king with a size several times larger than expected from the distribution of the rest of the populations of cities [17]. This king effect reveals a particular historical organization of France, whose roots are difficult to unravel. Nevertheless, we think that this king effect embodied by Paris is a significant signal to explain in order to understand the competition between cities in Europe.

### Kings and Crises in Complex Systems

We propose that these kings may reveal information which is complementary and perhaps sometimes even more important than the power law pdf.

Indeed, it is essential to realize that the long-term behavior of complex systems is often controlled in large part by rare catastrophic events: the universe was probably born during an extreme explosion (the “big-bang”); the nucleosynthesis of all important atomic elements constituting our matter results from the colossal explosion of supernovae; the largest earthquake in California repeating about once every two centuries accounts for a significant fraction of the state’s total tectonic deformation; landscapes are shaped more by the “millennium” flood that moves large boulders than by the action of all other eroding agents; the largest volcanic eruptions lead to major topographic changes as well as severe climatic disruptions; evolution is characterized by phases of quasi-stasis interrupted by episodic bursts of activity and destruction; financial crashes can destroy trillions of dollars in an instant; political crises and revolutions shape the long-term geopolitical landscape; even our personal life is shaped on the long run by a few key “decisions/happenstances.”

The outstanding scientific question is thus how such large-scale patterns of catastrophic nature might evolve from a series of interactions from the smallest to increasingly larger scales. In complex systems, it has been found that the organization of spatial and temporal correlations does not stem, in general, from a nucleation phase diffusing across the system. It results, rather, from a progressive and more global cooperative process occurring over

the whole system by repetitive interactions. An instance would be the many occurrences of simultaneous scientific and technical discoveries signaling the global nature of the maturing process.

Standard models and simulations of scenarios of extreme events are subject to numerous sources of error, each of which can have a negative impact on the validity of the predictions [93]. Some of the uncertainties are under control in the modeling process; they usually involve trade-offs between faithful descriptions and manageable calculations. Other sources of errors are beyond control as they are inherent in the modeling methodology of the specific disciplines. The two known strategies for modeling are both limited in this respect: analytical theoretical predictions are out of reach for most complex problems, while brute force numerical resolution of the equations (when they are known) or of scenarios is reliable only in the “center of the distribution”, i. e., in the regime far from the extremes where good statistics can be accumulated. Crises are extreme events that occur rarely, albeit with extraordinary impact, and are thus completely under-sampled and poorly constrained. Even the introduction of teraflop (or even petaflops in the near future) supercomputers does not qualitatively change this fundamental limitation.

Recent developments suggest that non-traditional approaches, based on the concepts and methods of statistical and nonlinear physics could provide a middle way to direct the numerical resolution of more realistic models and the identification of relevant signatures of impending catastrophes. Enriching the concept of self-organizing criticality, the predictability of crises would then rely on the fact that they are fundamentally outliers, e. g., large earthquakes are not scaled-up versions of small earthquakes but the result of specific collective amplifying mechanisms. To address this challenge, the available theoretical tools comprise in particular bifurcation and catastrophe theories, dynamical critical phenomena and the renormalization group, nonlinear dynamical systems, and the theory of partially (spontaneously or not) broken symmetries. Some encouraging results have been gathered on concrete problems, such as the prediction of the failure of complex engineering structures, the detection of precursors of stock market crashes and of human parturition, with exciting potential for earthquakes. At the beginning of the third millennium, it is tempting to extrapolate and forecast that a larger multidisciplinary integration of the physical sciences together with artificial intelligence and soft-computational techniques, fed by analogies and fertilization across the natural sciences, will provide a better understanding of the limits of predictability of catastro-

phes and adequate measures of risks for a more harmonious and sustainable future for our complex world.

### Future Directions

Our exposition has focused mainly on the concept of distributions of event sizes as a first approach to characterizing the organization of complex systems. But probability distribution functions are just one-point statistics and thus provide only an incomplete picture of the organization of complex systems. This opens the road to several better measures of the organization of complex systems.

- Statistical estimations of probability distribution functions is a delicate art. An active research field in mathematical statistics which is insufficiently used by practitioners of other sciences is the domain of “robust estimation.” Robust estimation techniques are methods which are insensitive to small departures from the idealized assumptions which have been used to optimize the algorithm. Such techniques include M-estimates (which follow from maximum likelihood considerations), L-estimates (which are linear combinations of order statistics), and R-estimates (based on statistical rank tests) [94,95,96].
- Ideally, one would like to measure the full multivariate distribution of events, which can be in full generality decomposed into the set of marginal distributions discussed above and of the copula of the system. A copula embodies completely the entire dependence structure of the system [97,98]. Copulas have recently become fashionable in financial mathematics and in financial engineering [19]. Their use in other fields in the natural sciences is embryonic but can be expected to blossom.
- When analyzing a complex system, a common trap is to assume without critical thinking and testing that the statistics are stationary, implying that monovariate (marginal) and multivariate distribution functions are sufficient to fully characterize the system. It is indeed a common experience that the dependencies estimated and predicted by standard models change dramatically at certain times. In other words, statistical properties are conditional on specific regimes. The existence of regime-dependent statistical properties has been discussed in particular in climate science, in medical sciences and in financial economics. In the latter, a quite common observation is that investment strategies, which have some moderate beta (coefficient of regression to the market) for normal times, can see their beta jump to a much larger value (close to 1 or larger depending on the leverage of the investment) at certain times when the market collectively dives. Said dif-

ferently, investments which are thought to be hedged against negative global market trends may actually lose as much or more than the global market at certain times when a large majority of stocks plunge simultaneously. In other words, the dependency structure and the resulting distributions at different time scales may change in certain regimes.

The general problem of the application of mathematical statistics to non-stationary data (including non-stationary time series) is very important, but alas, not much can be done. There are only a few approaches which may be used and only in specific conditions, which we briefly mention.

1. Use of algorithms and methods which are robust with respect to possible non-stationarity in data, such as normalization procedures or the use of quantile samples instead of initial samples.
  2. Modeling non-stationarity by some low-frequency random processes, such as a narrow-band random process  $X(t) = A(t) \cos(\omega t + \phi(t))$  where  $\omega \ll 1$  and  $A(t)$  and  $\phi(t)$  are slowly varying amplitude and phase. In this case, the Hilbert transform can be very useful to characterize  $\phi(t)$  non-parametrically.
  3. The estimation of the parameters of a low-frequency process based on a “short” realization is often hopeless. In this case, the only quantity which can be evaluated is the uncertainty (or scatter) of the results due to non-stationarity.
  4. Regime Switching popularized by Hamilton [99] for autoregressive time series models is a special case of non-stationarity, which can be handled with specific methods.
- We already discussed the problem of “kings.” One key issue that needs more scrutiny is that these outliers are often identified only with metrics adapted to take into account transient increases of the time dependence in the time series, as for instance in the case of returns of individual financial assets [88] (see also Chap. 3 of [91]). These outliers seem to belong to a statistical population which is different from the bulk of the distribution and require some additional amplification mechanisms active only at special times. The presence of such outliers both in marginal distributions and in concomitant events, together with the strong impact of crises and of crashes in complex systems, suggests the need for novel measures of dependence, different definitions of events and other time-varying metrics across different variables. This program is part of the more general need for a joint multi-time-scale and multi-variate approach to the statistics of complex systems.

- The presence of outliers poses the problem of exogeneity versus endogeneity. An event identified as anomalous could perhaps be cataloged as resulting from exogenous influences. The concept of exogeneity is fundamental in statistical estimation [100,101]. Here, we refer to the question of exogeneity versus endogeneity in the broader context of self-organized criticality, inspired in particular by the physical and natural sciences. As we already discussed, according to self-organized criticality, extreme events are seen to be endogenous, in contrast with previous prevailing views (see for instance the discussion in [64,102]). But, how can one assert with 100% confidence that a given extreme event is really due to an endogenous self-organization of the system, rather than a response to an external shock? Most natural and social systems are indeed continuously subjected to external stimulations, noises, shocks, solicitations, and forcing, which can vary widely in amplitude. It is thus not clear a priori if a given large event is due to a strong exogenous shock, to the internal dynamics of the system, or to a combination of both. Addressing this question is fundamental for understanding the relative importance of self-organization versus external forcing in complex systems and underpins much of the problem of dependence between variables. The concepts of endogeneity and exogeneity have many applications in the natural and social sciences (see [103] for a review) and we expect this viewpoint to develop into a general strategy of investigation.

## Bibliography

1. Satinover JB, Sornette D (2007) “Illusion of Control” in Minority and Parrondo Games. *Eur Phys J B* 60:369-384
2. Satinover JB, Sornette D (2007) Illusion of Control in a Brownian Game. *Physica A* 386:339-344
3. Sornette D (2006) Critical Phenomena in Natural Sciences. Chaos, Fractals, Self-organization and Disorder: Concepts and Tools, 2nd edn. Springer Series in Synergetics. Springer, Heidelberg
4. Feller W (1971) An Introduction to Probability Theory and its Applications, vol II. John Wiley, New York
5. Ruelle D (2004) Conversations on Nonequilibrium Physics With an Extraterrestrial. *Phys Today* 57(5):48-53
6. Zajdenweber D (1976) Hasard et Prévision. *Economica*, Paris
7. Zajdenweber D (1997) Scale invariance in Economics and Finance. In: Dubrulle B, Graner F, Sornette D (eds) *Scale Invariance and Beyond*. EDP Sciences and Springer, Berlin, pp 185-194
8. Malcai O, Lidar DA, Biham O, Avnir D (1997) Scaling range and cutoffs in empirical fractals. *Phys Rev E* 56:2817-2828
9. Biham O, Malcai O, Lidar DA, Avnir D (1998) Is nature fractal? *Response Sci* 279:785-786
10. Biham O, Malcai O, Lidar DA, Avnir D (1998) Fractality in nature. *Response Sci* 279:1615-1616



11. Mandelbrot BB (1998) Is nature fractal? *Sci* 279:783–784
12. Mandelbrot BB (1982) *The fractal Geometry of Nature*. Freeman WH, San Francisco
13. Aharony A, Feder J (eds) (1989) *Fractals in Physics*. Phys D 38(1–3). North Holland, Amsterdam
14. Riste T, Sherrington D (eds) (1991) *Spontaneous Formation of Space-Time Structures and Criticality*. Proc NATO ASI, Geilo, Norway. Kluwer, Dordrecht
15. Pisarenko VF, Sornette D (2003) Characterization of the frequency of extreme events by the Generalized Pareto Distribution. *Pure Appl Geophys* 160(12):2343–2364
16. Jessen AH, Mikosch T (2006) Regularly varying functions. *Publ l'Inst Math, Nouvelle serie* 79(93):1–23. Preprint [http://www.math.ku.dk/~mikosch/Preprint/Anders/jessen\\_mikosch.pdf](http://www.math.ku.dk/~mikosch/Preprint/Anders/jessen_mikosch.pdf)
17. Laherrère J, Sornette D (1999) Stretched exponential distributions in nature and economy: Fat tails with characteristic scales. *Eur Phys J B* 2:525–539
18. Malevergne Y, Pisarenko VF, Sornette D (2005) Empirical Distributions of Log>Returns: between the Stretched Exponential and the Power Law? *Quant Fin* 5(4):379–401
19. Malevergne Y, Sornette D (2006) *Extreme Financial Risks (From Dependence to Risk Management)*. Springer, Heidelberg
20. Frisch U, Sornette D (1997) Extreme deviations and applications. *J Phys I, France* 7:1155–1171
21. Willekens E (1988) The structure of the class of subexponential distributions. *Probab Theory Relat Fields* 77:567–581
22. Embrechts P, Klüppelberg CP, Mikosch T (1997) *modeling Extremal Events*. Springer, Berlin
23. Stuart A, Ord K (1994) *Kendall's advances theory of statistics*. John Wiley, New York
24. Bak P (1996) *How Nature Works: the Science of Self-organized Criticality*. Copernicus, New York
25. Sornette D (1994) Sweeping of an instability: an alternative to self-organized criticality to get power laws without parameter tuning. *J Phys I Fr* 4:209–221
26. Sornette D (2002) Mechanism for Power laws without Self-Organization. *Int J Mod Phys C* 13(2):133–136
27. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46:323–351
28. Wilson KG (1979) Problems in physics with many scales of length. *Sci Am* 241:158–179
29. Stauffer D, Aharony A (1994) *Introduction to Percolation Theory*, 2nd edn. Taylor & Francis, London, Bristol, PA
30. Gabrielov A, Newman WI, Knopoff L (1994) Lattice models of Fracture: Sensitivity to the Local Dynamics. *Phys Rev E* 50:188–197
31. Jensen HJ (2000) Self-Organized Criticality: Emergent Complex Behavior. In: *Physical and Biological Systems*. Cambridge Lecture Notes in Physics, Cambridge University Press
32. Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. *Internet Math* 1:226–251
33. Simkin MV, Roychowdhury VP (2006) Re-inventing Willis. Preprint <http://arXiv.org/abs/physics/0601192>
34. Sethna JP (2006) *Crackling Noise and Avalanches: Scaling, Critical Phenomena, and the Renormalization Group*. Lecture notes for Les Houches summer school on Complex Systems, summer 2006
35. Redner S (2001) *A Guide to First-Passage Processes*. Cambridge University Press, New York
36. Zygałło R (2006) Flashing annihilation term of a logistic kinetic as a mechanism leading to Pareto distributions. *Phys Rev E* 77, 021130
37. Stauffer D, Sornette D (1999) Self-Organized Percolation Model for Stock Market Fluctuations. *Phys A* 271(3–4):496–506
38. Kesten H (1973) Random difference equations and renewal theory for products of random matrices. *Acta Math* 131:207–248
39. Solomon S, Richmond P (2002) Stable power laws in variable economies. Lotka-Volterra implies Pareto-Zipf. *Eur Phys J B* 27:257–261
40. Saichev A, Malevergne A, Sornette D (2007) Zipf law from proportional growth with birth-death processes (working paper)
41. Reed WJ, Hughes BD (2002) From Gene Families and Genera to Incomes and Internet File Sizes: Why Power Laws are so Common in Nature. *Phys Rev E* 66:067103
42. Cabrera JL, Milton JG (2004) Human stick balancing: Tuning Lévy flights to improve balance control. *Chaos* 14(3):691–698
43. Eurich CW, Pawelzik K (2005) Optimal Control Yields Power Law Behavior. *Int Conf Artif Neural Netw* 2:365–370
44. Platt N, Spiegel EA, Tresser C (1993) On-off intermittency: A mechanism for bursting. *Phys Rev Lett* 70:279–282
45. Heagy JF, Platt N, Hammel SM (1994) Characterization of on-off intermittency. *Phys Rev E* 49:1140–1150
46. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data. Preprint <http://arxiv.org/abs/0706.1062>
47. Gouriéroux C, Monfort A (1994) Testing non-nested hypotheses. In: Engle RF, McFadden DL (eds) *Handbook of Econometrics*, Volume IV. Elsevier Science, pp 2583–2637
48. Cardy JL (1988) *Finite-Size Scaling*. North Holland, Amsterdam
49. Cranmer K (2001) Kernel estimation in high-energy physics. *Comput Phys Commun* 136(3):198–207
50. Sornette D, Knopoff L, Kagan YY, Vanneste C (1996) Rank-ordering statistics of extreme events: application to the distribution of large earthquakes. *J Geophys Res* 101:13883–13893
51. Pacheco JF, Scholz C, Sykes L (1992) Changes in frequency-size relationship from small to large earthquakes. *Nature* 355:71–73
52. Main I (2000) Apparent Breaks in Scaling in the Earthquake Cumulative Frequency-magnitude Distribution: Fact or Artifact? *Bull Seismol Soc Am* 90:86–97
53. Gabaix X, Ibragimov R (2008) Rank-1/2: A simple way to improve the OLS estimation on tail exponents. *Work Paper NBER*
54. Hill BM (1975) A simple general approach to inference about the tail of a distribution. *Ann Stat* 3:1163–1174
55. Drees H, de Haan L, Resnick SI (2000) How to Make a Hill Plot. *Ann Stat* 28(1):254–274
56. Resnik SI (1997) Discussion of the Danish Data on Large Fire Insurance Losses. *Astin Bull* 27(1):139–152
57. Pisarenko VF, Sornette D, Rodkin M (2004) A new approach to characterize deviations in the seismic energy distribution from the Gutenberg–Richter law. *Comput Seism* 35:138–159
58. Pisarenko VF, Sornette D (2006) New statistic for financial return distributions: power law or exponential? *Phys A* 366:387–400
59. Lasocki S (2001) Quantitative evidences of complexity of magnitude distribution in mining-induced seismicity: Implications for hazard evaluation. 5th International Symposium

- on Rockbursts and Seismicity in Mines. In: van Aswegen G, Durrheim RJ, Ortlepp WD (eds) *Dynamic Rock Mass Response to Mining*, Symp Ser, vol 527, S Afr Inst Min Metall, Johannesburg, pp 543–550
60. Lasocki S, Papadimitriou EE (2006) Magnitude distribution complexity revealed in seismicity from Greece. *J Geophys Res* B11309(111). doi:10.1029/2005JB003794
  61. Anderson PW (1972) More is different (Broken symmetry and the nature of the hierarchical structure of science). *Science* 177:393–396
  62. Sornette D, Davis AB, Ide K, Vixie KR, Pisarenko VF, Kamm JR (2007) Algorithm for Model Validation: Theory and Applications. *Proc Nat Acad Sci USA* 104(16):6562–6567
  63. Geller RG, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. *Science* 275(5306):1616–1617
  64. Bak P, Paczuski M (1995) Complexity, contingency and criticality. *Proc Nat Acad Sci USA* 92:6689–6696
  65. Allègre CJ, Le Mouél JL, Provost A (1982) Scaling rules in rock fracture and possible implications for earthquake predictions. *Nature* 297:47–49
  66. Keilis-Borok V (1990) The lithosphere of the Earth as a large nonlinear system. *Geophys Monogr Ser* 60:81–84
  67. Sornette A, Sornette D (1990) Earthquake rupture as a critical point: Consequences for telluric precursors. *Tectonophysics* 179:327–334
  68. Bowman DD, Ouillon G, Sammis CG, Sornette A, Sornette D (1996) An observational test of the critical earthquake concept. *J Geophys Res* 103:24359–24372
  69. Sammis SG, Sornette D (2002) Positive Feedback, Memory and the Predictability of Earthquakes. *Proc Nat Acad Sci USA (SUPP1)* 99:2501–2508
  70. Huang Y, Saleur H, Sammis CG, Sornette D (1998) Precursors, aftershocks, criticality and self-organized criticality. *Europhys Lett* 41:43–48
  71. National Institute of Standards and Technology (2007) *Engineering Statistical Handbook*, National Institute of Standards and Technology. Preprint <http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>
  72. Pollock AA (1989) Acoustic Emission Inspection. In: *Metal Handbook*, 9th edn, vol 17, Nondestructive Evaluation and Quality Control. ASM International, pp 278–294
  73. Omeltchenko A, Yu J, Kalia RK, Vashishta P (1997) Crack Front Propagation and Fracture in a Graphite Sheet: a Molecular-dynamics Study on Parallel Computers. *Phys Rev Lett* 78:2148–2151
  74. Fineberg J, Marder M (1999) Instability in Dynamic Fracture. *Phys Rep* 313:2–108
  75. Lei X, Kusunose K, Rao MVMS, Nishizawa O, Sato T (2000) Quasi-static Fault Growth and Cracking in Homogeneous Brittle Rock Under Triaxial Compression Using Acoustic Emission Monitoring. *J Geophys Res* 105:6127–6139
  76. Wesnousky SG (1994) The Gutenberg-Richter or characteristic earthquake distribution, which is it? *Bull Seismol Soc Am* 84(6):1940–1959
  77. Wesnousky SG (1996) Reply to Yan Kagan's comment On: The Gutenberg-Richter or characteristic earthquake distribution, which is it? *Bull Seismol Soc Am* 86(1A):286–291
  78. Kagan YY (1993) Statistics of characteristic earthquakes. *Bull Seismol Soc Am* 83(1):7–24
  79. Kagan YY (1996) Comment On: The Gutenberg-Richter or characteristic earthquake distribution, which is it? by Wesnousky SG. *Bull Seismol Soc Am* 86:274–285
  80. Gil G, Sornette D (1996) Landau-Ginzburg theory of self-organized criticality. *Phys Rev Lett* 76:3991–3994
  81. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of Earthquakes in Simple Models of Heterogeneous Faults. *Phys Rev Lett* 78:4885–4888
  82. Ben-Zion Y, Eneva M, Liu Y (2003) Large Earthquake Cycles And Intermittent Criticality On Heterogeneous Faults Due To Evolving Stress And Seismicity. *J Geophys Res* B6(108):2307. doi:10.1029/2002JB002121
  83. Hillers G, Mai PM, Ben-Zion Y, Ampuero J-P (2007) Statistical Properties of Seismicity Along Fault Zones at Different Evolutionary Stages. *Geophys J Int* 169:515–533
  84. Zöller G, Ben-Zion Y, Holschneider M (2007) Estimating recurrence times and seismic hazard of large earthquakes on an individual fault. *Geophys J Int* 170:1300–1310
  85. Ben-Zion (2007) private communication
  86. L'vov VS, Pomyalov A, Procaccia I (2001) Outliers, Extreme Events and Multiscaling. *Phys Rev E* 6305(5):6118, U158-U166
  87. Johansen A, Sornette D (1998) Stock market crashes are outliers. *Eur Phys J B* 1:141–143
  88. Johansen A, Sornette D (2001) Large Stock Market Price Drawdowns Are Outliers. *J Risk* 4(2):69–110. <http://arXiv.org/abs/cond-mat/0010050>
  89. Johansen A, Sornette D (2007) Shocks, Crash and Bubbles in Financial Markets. In press, In: *Brussels Economic Review on Non-linear Financial Analysis* 149–2/Summer 2007. Preprint <http://arXiv.org/abs/cond-mat/0210509>
  90. Gopikrishnan P, Meyer M, Amaral LAN, Stanley HE (1998) Inverse cubic law for the distribution of stock price variations. *Eur Phys J B* 3:139–140
  91. Sornette D (2003) *Why Stock Markets Crash, Critical Events in Complex Financial Systems*. Princeton University Press, Princeton, NJ
  92. Zipf GK (1949) *Human behavior and the principle of least-effort*. Addison-Wesley, Cambridge, MA
  93. Karplus WJ (1992) *The Heavens are Falling: The Scientific Prediction of Catastrophes in Our Time*. Plenum, New York
  94. Hubert PJ (2003) *Robust Statistics*. Wiley-Interscience, New York
  95. van der Vaart AW, Gill R, Ripley BD, Ross S, Silverman B, Stein M (2000) *Asymptotic Statistics*. Cambridge University Press, Cambridge
  96. Wilcox RR (2004) *Introduction to Robust Estimation and Hypothesis Testing*, 2nd edn. Academic Press, Boston
  97. Joe H (1997) *Multivariate models and dependence concepts*. Chapman & Hall, London
  98. Nelsen RB (1998) *An Introduction to Copulas*, Lectures Notes in statistic 139. Springer, New York
  99. Hamilton JD (1989) A New Approach to the Economic Analysis of Non-stationary Time Series and the Business Cycle. *Econometrica* 57:357–384
  100. Engle RF, Hendry DF, J Richard F (1983) Exogeneity. *Econometrica* 51:277–304
  101. Ericsson N, Irons JS (1994) *Testing exogeneity*, Advanced Texts in Econometrics. Oxford University Press, Oxford
  102. Sornette D (2002) Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth. *Proc Nat Acad Sci USA* 99(SUPP1):2522–2529

103. Sornette D (2005) Endogenous versus exogenous origins of crises, in the monograph entitled: *Extreme Events in Nature and Society*. In: Albeverio S, Jentsch V, Kantz H (eds) *Series: The Frontiers Collection*. Springer, Heidelberg (e-print at <http://arxiv.org/abs/physics/0412026>)

## Probability and Statistics in Complex Systems, Introduction to

HENRIK JELDTTOFT JENSEN<sup>1,2</sup>

<sup>1</sup> Institute for Mathematical Sciences, London, UK

<sup>2</sup> Department of Mathematics, Imperial College London, London, UK

Is the nature of complex systems such that they are in need of a special treatment in terms of statistics and probability. Yes, they are in the sense that a particular focus suggests itself. This becomes clear if we try to specify what we mean by complex systems. Although no consensus exists for the definition of what constitutes a complex system, the following summary will probably be accepted by most people.

- **Complex Systems** consist of a large number of interacting components. The interactions give rise to emergent hierarchical structures. The components of the system and properties at systems level typically change with time. A complex system is inherently open and its boundaries often a matter of convention.

Large numbers of components and interaction between these are central and this has immediate consequences for the nature of the relevant statistics. The interactions between the many components will typically be so strong that correlations cannot be neglected (► [Correlations in Complex Systems](#)) and hence, when we sum up contributions from the individual parts, we may not have the central limit theorem to ensure that the macroscopic quantities are Gaussian distributed. Instead of peaked distributions with well-defined average and higher moments, one typically encounters very broad heavy tailed distributions, frequently the tail reaches so far out that the average doesn't even exist. When this is the case, a description in terms of the typical—on the average scenario—is not possible (► [Probability Distributions in Complex Systems](#)). This makes not only fluctuations significant, it makes a description and understanding of the fluctuations absolutely essential (► [Fluctuations, Importance of: Complexity in the View of Stochastic Processes](#)). To illustrate this point think of, say, flood protection. It is no good to protect oneself against some imaginary average event, say a “typical” flooding, if effectively any size of flooding may occur with a non negligible probability. This makes it important to be

able to deal with broad distributions and distributions of extreme events (► [Extreme Value Statistics](#)).

The hierarchical nature of complex systems (► [Hierarchical Dynamics](#)) leads to situations where the existence of many time scales cannot be ignored and as a result it may be inappropriate to assume the statistics to be stationary. The time dependence of the statistics may come about for a number of reasons. One is that the emergent hierarchical components change their internal state with time. An alternative reason may be that the strong interaction between components, themselves with no internal time evolution, prevents the collective set of components from reaching an equilibrium or asymptotic stationary state. In such cases one needs to be able to understand how to describe and predict the behavior of a system that is in a transient for all the relevant time scales, such as, say, the age of life on earth, if one is dealing with the history of the biosphere.

It seems natural to divide the discussion of probability and statistics for complex systems into at least three aspects:

1. Analyzing data from complex systems,
2. The phenomenology and
3. Modeling.

When analyzing data it is important to be able to handle the effect of long memory, correlations and exceptionally strong fluctuations. These effects make it necessary to exert special care when trying to identify probability distributions extracted from observational or experimental data or from data generated by computer simulations (► [Probability Densities in Complex Systems, Measuring](#)). The often very large amounts of data, and the lack of fundamental theory derived from first principle, have made it important to develop methods to identify structures in data sets, in this respect Bayesian statistics is often very relevant to complex systems (► [Bayesian Statistics](#)).

From experiments and observations we know that one signature of complexity may be power law like probability distributions. To determine the exponent can be complicated as the exponent one reads off from the slope in a log-log plot of the distribution may only be an “apparent” exponent. Not that this makes the exponent less important, but it does make the exponent more specific. An apparent exponent might very well depend on systems size and on the amount of collected data. In contrast we know from the lesson of statistical mechanics of critical phenomena that the asymptotic behavior of an infinite system may be the same even when microscopic details differ. This encourages the study of simplistic models in the hope that, though simplistic, the models might nevertheless capture

the essential mechanisms relevant for the phenomena under consideration. With this in mind, particular emphasis is placed on how apparent exponents converge towards their “universal” values in the limit of long distance and long time scales in the limit of a system composed of infinitely many components. Of course we don’t know how applicable this form of universality is in the case of non-equilibrium complex systems. It might very well be that the many power laws observed in experiments are in fact only approximate and only persist for a limited range of support for the stochastic variable being studied. To settle this, careful experiments and theoretical studies are needed and will make use of some of the methods relevant to power laws discussed in the present section of the encyclopedia.

The focus on the functional form of the probability distributions becomes complicated by the fact that many complex systems never enter a stationary state during time scales accessible to observations or simulations. This doesn’t exclude that the system relaxes towards a stationary state in the mathematically asymptotic limit of infinite time, but this limit might be of little physical relevance. Then it becomes essential to study the very nature of how systems for all observational times is relaxing towards the stationary state. Inspired by the slow dynamics in glassy systems, and the very widespread intermittent dynamics encountered when many components interact, it has been suggested that the statistics of records may be of relevance to the intermittent relaxation of complex systems. The analysis of the statistics of the time instances marked by the occurrence of abrupt activity can be interesting. It might allow insight into the question concerning, how relevant record dynamics is for the complex dynamics and can in this way help to provide understanding of the collective dynamics of the components ([► Record Statistics and Dynamics](#)).

To construct and analyze theoretical models of complex systems a number of methods have been developed. At the intuitive level we have attempts to develop phenomenological theories by generalizing the well-studied branching process originating in Galton’s and Watson’s sociological studies ([► Branching Processes](#)). Among attempts to formulate theory of more basic foundation we encounter methods developed in physics to deal with phase transitions in materials where many interacting particles enter into a critical state. Here a critical state is taken to mean that no particular length or time scale can be identified, one talks about scale invariance, to indicate that all length and time scales are involved in the phenomena. To analyze such systems field theoretic methods are particularly useful as a tool to study the asymptotic behavior,

i. e., long length and time scales ([► Field Theoretic Methods](#)). Stochastic analysis is particularly relevant since we are dealing with large numbers of components and, moreover, a stochastic element is often present even at the basic level, e. g. through thermal fluctuations in the case of physical systems ([► Stochastic Processes](#)).

In traditional statistical mechanics the concept of entropy has played a very important role. It is therefore only natural that work is being done to try to develop the definition of entropy to make it applicable beyond the traditional areas. In particular strong interaction makes the entropy non-extensive and appropriate generalizations are needed ([► Entropy in Ergodic Theory](#)).

The ordinary random walker is a bit of a workhorse in statistical mechanics. When random walks are used to model aspects of complex systems, the walker needs often to be dressed up with a broad distribution of step sizes ([► Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion](#)) or to walk in a background that makes the step size distribution location dependent ([► Random Walks in Random Environment](#)). This more sophisticated walker is not any longer necessarily controlled by the central limit theorem and new mathematics are needed.

The theory of random matrices is another example of methods developed to understand statistical aspects of physical systems that now have become of much broader importance. This is a field developed in response to the need of physicists when they, a while ago, started to try to understand the energy spectrum of heavy nuclei. The formalism has since then been used in a range of very different fields including the analysis of the stability of ecosystems. It is very likely that random matrices also will become a standard tool in complexity ([► Random Matrix Theory](#)). One reason for this is that complex systems often can be represented in terms of networks and that random matrix theory naturally relates to network analysis.

Finally, I feel it is in place to explain why this section contains an article about the stochastic Löwner equation ([► Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems](#)). The topic is included, as a spectacular example of how successful fairly abstract mathematics, from the realms of the pure end of the mathematical spectrum, can sometimes be in providing a detailed quantitative understanding of the intricacies of systems that are complex in the sense that they are far from equilibrium and in a non-stationary state. One might hope that this degree of detail in the mathematical understanding of the statistics of complex systems may become the norm in the future, as the research in to the statistical and probabilistic analysis of complex systems develop.



## Protein Mechanics at the Single-Molecule Level

MARIANO CARRIÓN-VÁZQUEZ<sup>1</sup>, MAREK CIEPLAK<sup>2</sup>,  
ANDRÉS F. OBERHAUSER<sup>3</sup>

<sup>1</sup> Cajal Institute, CSIC & CIBERNED (Network on Degenerative Diseases), Madrid, Spain

<sup>2</sup> Institute of Physics, Polish Academy of Sciences, Warsaw, Poland

<sup>3</sup> Department of Neuroscience and Cell Biology, Department of Biochemistry and Molecular Biology and Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Background](#)

[Methodological Bases](#)

[Main Findings](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Atomic force microscopy (AFM)** A near-field type of microscopy that uses a mechanical force sensor (*cantilever*) and a nanopositioner to either scan a surface, obtaining a *topography* of the sample (“imaging” configuration) and/or to stretch the sample to obtain a *force spectrum* (“force spectroscopy” configuration). This acronym is also used throughout the text for the instrument itself, the atomic force microscope.

**Bionanomachines** Nanoscopic structures, usually constituted by protein complexes, which perform most cell functions with a few exceptions like genetic information storage.

**Enthalpic elasticity** Elasticity derived from the breaking of non-covalent bonds after stretching.

**Entropic elasticity** In polymer science, this type of elasticity results from the entropic behavior of monomers in a polymer after its stretching. This recoiling is driven by thermal energy (second principle of thermodynamics) and results in the so-called *restoring force*.

**Functionalization** Process by which specific chemical groups are added typically to a surface and/or the sample to gain certain control of sample immobilization (attachment, orientation, coverage, etc.)

**Hookean spring** A spring that shows a linear force-extension relationship as is the case of an AFM cantilever.

**Mechanical stability** It can be operationally defined as the amplitude of the highest force peak ( $F_{\max}$ ) observed during the stretching of a single protein molecule using the length-clamp mode of SMFS, averaged over many unraveling events. Most proteins show just one force peak.

**Molecular chaperons** Protein complexes that help protein folding and complex assembly. Chaperonins are a well characterized subclass of chaperons that assist in the folding of newly-made proteins in all cells. These bionanomachines use chemical energy, in the form of adenosine triphosphate (ATP).

**Molecular dynamics** A specialized discipline of *molecular modeling* and *computer simulation* based on statistical mechanics. These techniques are used to simulate the behavior of molecules from the physicochemical principles.

**Polyprotein** In protein engineering, artificial polymeric protein formed by repeats (oriented or not) of a protein or a protein domain linked by covalent (*peptide*, *isopeptide* or *disulfide*) bonds. Polymerization can be achieved at the DNA (by genetic engineering techniques: *in vivo*) or protein (by using biochemical techniques: *in vitro*) level.

**Protein** Natural biopolymer composed of up to 20 different monomers, *amino acids*, linked by so-called *peptide bonds* (a planar covalent bond), which typically acquires a unique 3D (*fold*) structure. The sequence of amino acids of a polypeptide is its *primary structure*. Proteins with *quaternary* structure are formed by several polypeptides, which are linked by non-covalent bonds, other covalent bonds, or both.

**Protein engineering** Discipline at the crossroads of molecular biology (includes genetic engineering, a technology to manipulate DNA), biochemistry, structural biology, and bioinformatics that aims to either the modification of proteins to improve or study its properties (redesign), or to design new proteins *de novo*. Usually it involves making changes in the sequence of a gene coding for a protein (usually by *polymerase chain reaction* and *directed mutagenesis*) in order to bring about desirable changes in its structure and/or function.

**Protein folding** Process by which a polypeptide acquires its native 3D structure (*fold*).

**Protein fold** Unique 3D structure of proteins (also called *superior structure* or *protein conformation*), achieved by either self-assembly alone or with the help of specific proteins (*molecular chaperons*), which is neces-



sary for its biological function. There are several structural levels of protein conformation: *secondary* (the main models are  $\alpha$ -helix and  $\beta$ -sheet; the latter formed by  $\beta$ -strands), *tertiary* (final fold of a polypeptide; e. g.  $\beta$ -sandwich  $\beta$ -barrel), *quaternary* (resulting from the association of several polypeptides usually by non-covalent bonds). Protein structures (folds) resolved at atomic resolution are unique having an ascribed file specifying their atomic coordinates (i. e., Protein Data Bank, PDB file) and are classified somewhat artificially into discrete classes (e. g., immunoglobulin fold).

**Protein nanomechanics** New discipline in charge of measuring forces, distances, motions, energies, and deformations involved in the manipulation of individual proteins or protein complexes, which are typically in the sub-micrometer and sub-nanonewton ranges.

**Protein unfolding** Process by which a native protein loses its native folding becoming “*denatured*”.

**Proteasomes** Large bionanomachines that degrade damaged or unneeded proteins by *proteolysis* (a chemical reaction that breaks peptide bonds). They are present in all kinds of cells, belonging to the class of enzymes called *proteases*, and they use ATP as a source of chemical energy.

**Single-molecule force spectroscopy (SMFS)**

Technique carried out by several instruments (*AFM*, *optical tweezers* or *biomembrane force probe*) consisting in stretching single molecules to measure the resistance forces (*length-clamp* mode) or distances traveled between resistance barriers (*force-clamp* mode).

**Unfolding (folding) pathway** Energetic representation of the pathway of an unfolding (folding) reaction.

## Definition of the Subject

Proteins can be considered as machine-like devices that function through complex structural changes in their intra- or intermolecular bonding. Understanding the dynamics of the inner workings of proteins is still one of the major challenges in biology.

Many proteins are nanomachines that use mechanical forces to fulfill a variety of cellular functions from replication to cell adhesion to cell crawling. The nanomachinery involved in these processes (i. e. the internal parts of these bionanomachines) is still poorly understood. Protein mechanics has emerged as a new multidisciplinary field to directly apply and measure mechanical forces through an array of recently developed dynamic techniques for manipulating single molecules, both in real time and under physiological conditions. After a decade, this field is still

maturing fast and exciting developments await just around the corner.

AFM (atomic force microscopy) single-molecule force spectroscopy (SMFS) is one of the main technical pillars of this new discipline and it is particularly suited to directly quantify the forces involved in both intra- and intermolecular protein interactions. In combination with protein engineering and computer simulations, this technique has been used to characterize the unfolding and refolding reactions in a variety of protein structures, both with and without “mechanical functions”.

Protein engineering allows the unequivocal identification of single molecules, through polypeptide analysis, and a careful experimental dissection of the experimental variables involved, through mutational analysis. Computer simulations based on molecular dynamics allow the process of the mechanical unfolding/folding of a protein to be modeled at the atomic level in order to obtain detailed structural information on its dynamics. These descriptions have been particularly useful in predicting and understanding the complexity behind the experimental results obtained by SMFS.

This review summarizes the concepts underpinning this field and some of the main findings to date.

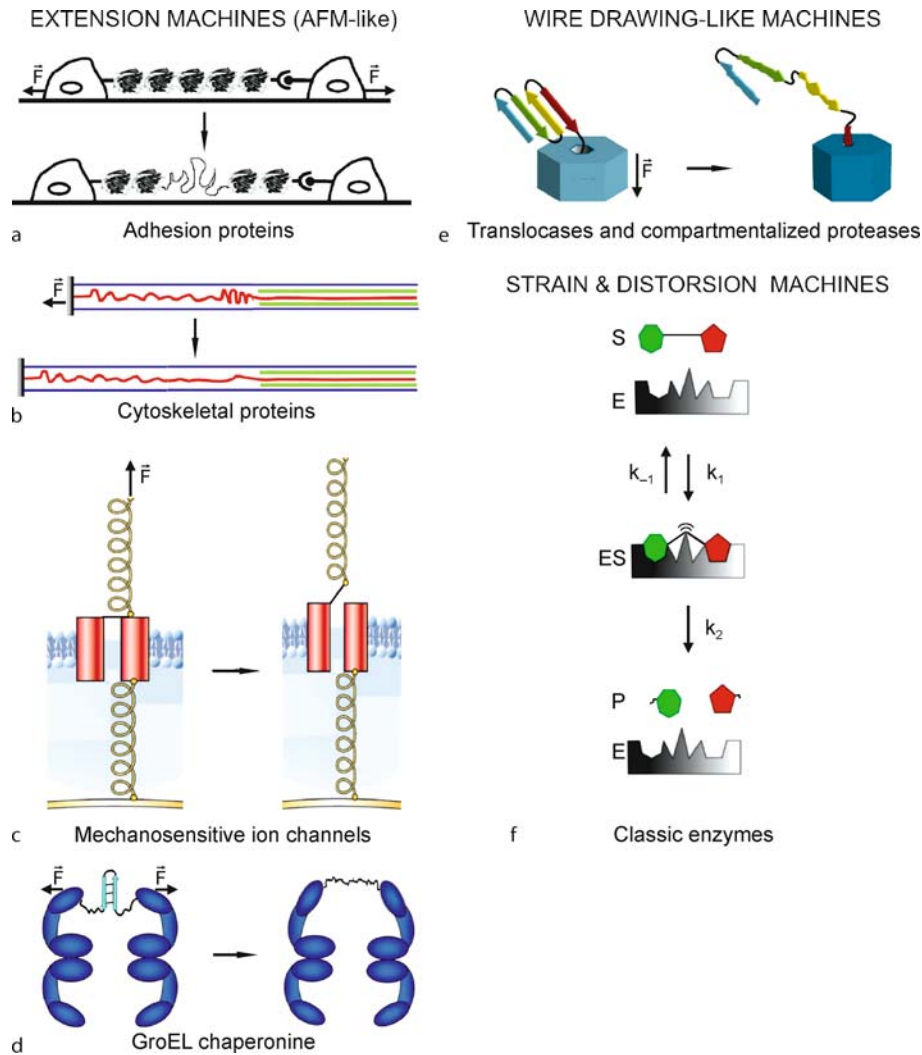
## Background

### Mechanical Force: A New Biochemical Parameter

In contemporary biology we tend to regard the cell as a factory-like system crowded with a variety of specialized molecular “machines” of nanometer size (i. e., nanomachines), mainly proteins that exist either as single polypeptides or as complex assemblies of protein “parts” [5]. Understanding the inner workings of these biological machines remains one of the more active frontiers in biology.

The function of an individual protein or protein complex often involves the conversion of chemical energy (stored or supplied) into mechanical work through conformational changes. These “bionanomachines” that make use of mechanical forces are located throughout the cell (from the cell nucleus to the extracellular matrix) and they are involved in processes as diverse as replication, transcription, translation, protein folding, protein and nucleic acid unfolding, protein degradation, nucleic acid and protein translocation, organelle transport, muscle elasticity, cell adhesion, membrane fusion, or cell crawling [19,53] (Fig. 1).

Mechanical forces are also crucial to regulate the structure and function of cells and tissues. Thus, the shape of eukaryotic cells, and by extension that of the multicellular organisms they form, is the result of mechanosen-



**Protein Mechanics at the Single-Molecule Level, Figure 1**

Mechanical biomachines. **a** Modular cell adhesion proteins may function as shock absorbers by increasing the range and lifetime of adhesion bonds. **b** Cytoskeletal proteins such as titin (in red, myosin filament in green, actin filaments in blue) may function as adjustable elastic springs. **c** Mechanosensitive ion channels are present in many biological systems some of which, like the auditory system have not yet been identified. The gating spring is a critical proteinaceous component of this machinery. **d** Chaperonins such as GroEL may induce conformational changes to mechanically unfold the substrate protein before refolding. **e** Compartmental proteases: the AAA+ (hexameric ring) ATPase from the proteasome and other related proteases unfold proteins, presumably by force, in an ATP-dependent manner, prior to their translocation to the catalytic chamber for degradation. **f** The strain and distortion hypothesis by Haldane and Pauling postulates that enzyme catalysis may work by inducing mechanical tension in the enzyme-substrate (ES) complex. (Modified from [27])

sory, mechanotransduction and mechanoreponse cycles. Responses to mechanical forces also underlie many biological processes from normal morphogenesis to carcinogenesis, cardiac hypertrophy, wound healing, and bone homeostasis. Indeed, recent studies show that several signaling pathways are involved in force transduction, including MAP kinases, small GTPases, and tyrosine kinases/phosphatases [6,48,49,121,122].

The molecular mechanisms by which mechanical forces influence these processes have been elusive due to the lack of appropriate tools. However, with the recent advent of single-molecule manipulation techniques we can now investigate these new biochemical pathways by directly probing bond dynamics in real time and under physiological conditions. These new techniques allow us to use mechanical force as an additional parameter in a biochem-

ical reaction, which can dramatically affect its rates in both directions.

### Single Molecule Biology: A New Scientific Revolution

Single-molecule biology (alternatively referred to as single-molecule, biophysics or biochemistry) is an entirely new field of science, at the crossroads of several disciplines (namely biology, physics, chemistry, material science, and computer science), which overcomes the restrictions of the traditional bulk biochemical studies by focusing not on a population of molecules but on the molecule itself. Single-molecule methods may be considered a “paradigm shift” as they allow the behavior of individual molecules to be analyzed directly (under thermodynamic equilibrium or nonequilibrium conditions) at their real “nanoscale” and in their own “nanoworld” where thermal motion is a dominant force (see below).

Single-molecule techniques allow us to test the so-called ergodic principle (see ► [Ergodic Theorems](#)) of molecular populations: the measurement of a property in an ensemble at a given time should be equivalent to the average of the property measured on a single molecule over long periods of time. Moreover, these methods are revealing important information regarding:

- a) Intermolecular variations in the experimental parameter of interest, which can arise from chemical (e.g. extent of glycosylation) or non-chemical (often called “static disorder”) differences,
- b) Rare events hidden in the ensemble,
- c) The distribution underlying the ensemble average,
- d) Fluctuations over time (the so-called “dynamic disorder”), and
- e) Molecule kinetics without requirement for synchronization [61,96,134].

An additional technical advantage of single-molecule techniques is that the data can be directly compared to that obtained *in silico* by molecular dynamics simulations, as these methods also deal with single molecules.

There are two main types of single-molecule methods: 1) those that do not use an external force such as single-molecule fluorescence microscopy [61,96,134]; and 2) those imposing an external force to the system through an electric field (e.g. patch-clamp [83]) or a mechanical manipulation (through tension or torsion). The latter subtype, the so-called single-molecule manipulation techniques, offers a unique opportunity to study the behavior of molecules under an external mechanical force, applied either directly using flexible beams (AFM, microneedles,

optical fibers) and vesicle membranes (biomembrane force probe), or through external-field manipulators (optical tweezers, magnetic tweezers, flow-field apparatus) [20,84]. These techniques have been used to examine the nanomechanics of the main biopolymers (DNA, RNA, polysaccharides, and proteins) [61,96,134]. Proteins are in charge of virtually every process that occurs in modern cells and are the subject of this review.

### Protein Nanomechanics

Nanomanipulation techniques can be used to study all kinds of proteins, although pioneering experiments concentrated for obvious reasons (namely function, modularity and size) on the so-called “mechanical proteins”. These are proteins with a mechanical function that can generate, transmit or use mechanical forces during their normal activity in the organism. They fall into two main subclasses: proteins that generate mechanical forces (biomolecular motors, which have been mainly probed by optical tweezers) and proteins that are subjected to the mechanical forces generated by biomolecular motors or from the environment (mainly probed by AFM).

In order to identify the molecular mechanisms involved in the activity of both types of proteins it is important to analyze their mechanics at the single-molecule level. “Protein nanomechanics” achieves this by studying the forces, distances, motions, energies, and deformations involved in individual proteins or protein complexes, typically in the sub-micrometer and sub-nanonewton ranges.

Mechanical forces have also been used to probe the mechanical strength of “non-mechanical proteins” (i.e., those with no known mechanical function) to better understand the thermodynamics and kinetics of protein unfolding/refolding. Moreover, intermolecular interactions in protein complexes, some of which are subjected to mechanical force *in vivo* (e.g. adhesion proteins), have also been studied by mechanical stretching.

We shall focus on the response of individual proteins to the mechanical forces applied mainly by AFM and will provide an overview highlighting the principles, ideas, achievements, and perspectives of this fast-growing multidisciplinary field. For further reading and details, the reader is recommended to consult other reviews on protein nanomechanics from the general ones, which include those focusing on molecular biomotors [19,34,53,79], to others restricted to specific types of interactions: inter- and intramolecular [27,80,133], intermolecular [51,66,125], intramolecular [44,86,98,132], and molecular interactions of native membrane proteins [60,82].

**Protein Mechanics at the Single-Molecule Level, Table 1**

Relevant forces in protein nanomechanics. Ubiquitous thermal forces help to overcome the energy activation barriers of biochemical reactions and are the basis for the entropic elasticity displayed by many proteins in solution, which is on the range of a few pN. Breaking the non-covalent bonds (van der Waals, hydrogen, electrostatic) that maintain protein folds and protein interactions needs higher forces, typically below 300 pN, while the rupture of covalent bonds requires stronger forces, above 1000 pN. Forces accessible to the AFM techniques are highlighted in *italics*

Force type	Force range (pN)	Origin	Biological role	Protein examples
Langevin (thermal agitation)	~0.001	thermal energy	activation of energy barrier of reactions	typical enzymatic reactions
<i>entropic</i>	~0–10	thermal energy (on a polymer)	entropic elasticity (recoil) of biopolymers	titin passive elasticity (physiological range)
<i>enthalpic (non covalent)</i>	~10–300	non -covalent bonds	folding/interactions in biopolymers	unfolding/unbinding in proteins
<i>enthalpic (covalent)</i>	~1000–3000	covalent bonds	biopolymer synthesis	Proteolysis (enzymatic hydrolysis of proteins)

**Methodological Bases**

**Range of Relevant Mechanical Forces in Biology**

What are the magnitudes of the biologically relevant forces that affect protein structure? Proteins are subject to thermal forces, which are random in nature. These forces are in the femtonewton ( $fN = 10^{-15}$  N) range and when they act on small objects like bionanomachines in solution, they result in what is called Brownian motion (Table 1). It is through thermal energy that proteins reach the high-energy transition states that are essential in biochemical reactions. To understand life at a fundamental level, it is important to know how these protein machines move their parts and change shape in response to the thermal and mechanical forces present in their nanoenvironment. The energies involved in protein conformational changes (the “signals” of our experiments) are just above thermal energy levels (“noise”), typically ranging from  $1 k_B T$  (thermal energy;  $k_B T = 4.1$  pNnm = 0.6 kcal/mol, at room temperature; where  $k_B$  is the Boltzmann constant and  $T$  is the absolute temperature) to  $25 k_B T$  (the energy released by ATP hydrolysis) such that the structures are stable enough to prevail at physiological temperatures. Given that changes in protein conformation are measured in the Ångstrom-nanometer range (Å-nm,  $1 \text{ Å} = 10^{-10}$  m,  $1 \text{ nm} = 10^{-9}$  m) the relevant biological forces are expected to be in the piconewton range ( $1 \text{ pN} = 10^{-12}$  N).

Because proteins are subject to thermal forces, the number of possible conformations (entropy) reaches a maximum when a protein forms a random coil or is denatured. Conformational entropy becomes progressively reduced with the formation of secondary and tertiary structures. Stretching random-coiled proteins in the low force regime to overcome “entropic forces” requires the application of forces in the order of a few pN, which has

been achieved experimentally using single-molecule manipulation techniques. Several molecular motors such as myosin, kinesin, and RNA or DNA polymerases also generate forces in this range.

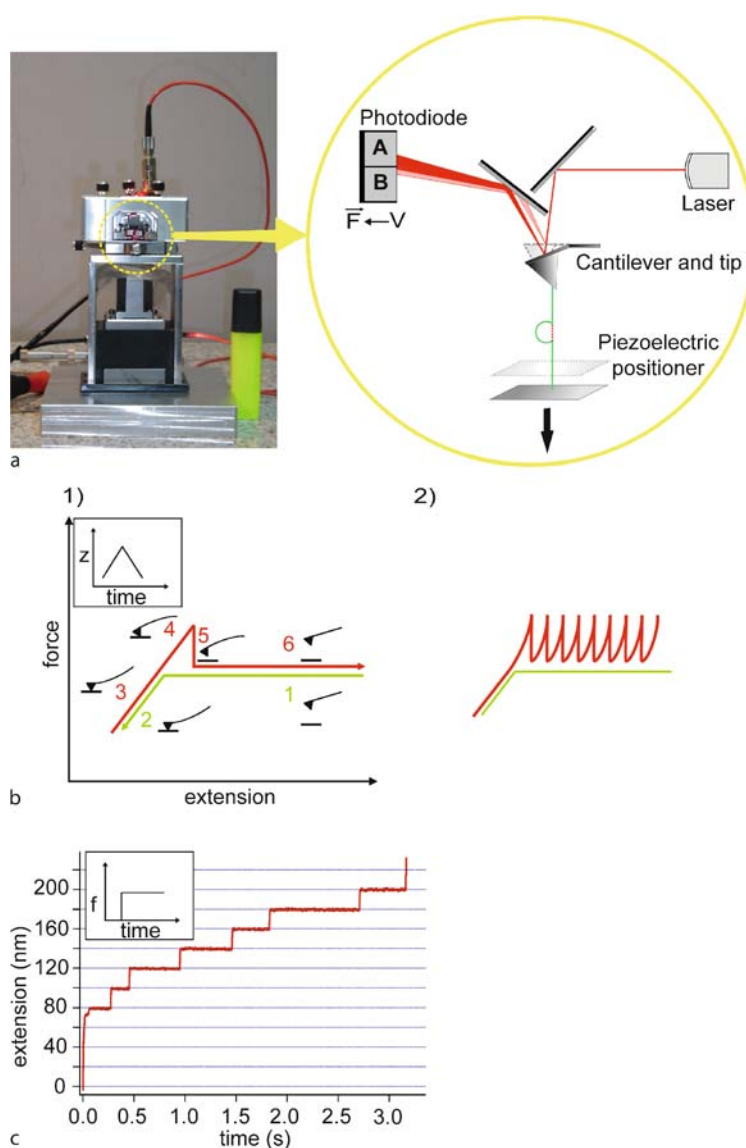
The next group are the “enthalpic forces”, which includes the forces needed to unfold the folded domains of proteins (i. e. intramolecular interactions) as well as those required to overcome specific intermolecular interactions such as ligand/receptor or antigen/antibody. These forces are typically in the 50–300 pN range, when measured at a high loading rate. It must be noted that protein mechanical unfolding is typically a non-equilibrium dynamic process and therefore, these forces depend on the loading rate (see definition below). The typical loading rates in vivo may in some cases be much lower and accordingly the corresponding forces may also be lower.

Finally, the forces needed to break covalent bonds apart are almost two orders of magnitude larger, in the range of a few nanonewtons ( $1 \text{ nN} = 10^{-9}$  N) [134].

**Single-Molecule Force Spectroscopy of Proteins: The Underlying Principle and Modes**

AFM is the main technique used to characterize the mechanical resistance of both individual polypeptides (intramolecular interactions) and protein-biomolecule bonds (intermolecular interactions) [20]. Thus most of studies on single protein stretching have used this technique [27] with the exception of three studies performed with optical tweezers [8,28,62]. The mechanical resistance of intermolecular interactions in protein pairs have also been studied using the biomembrane force probe [41].

The AFM was originally developed as a high-resolution imaging tool [15] before it began to be used to probe and manipulate atoms and molecules. The so-called



### Protein Mechanics at the Single-Molecule Level, Figure 2

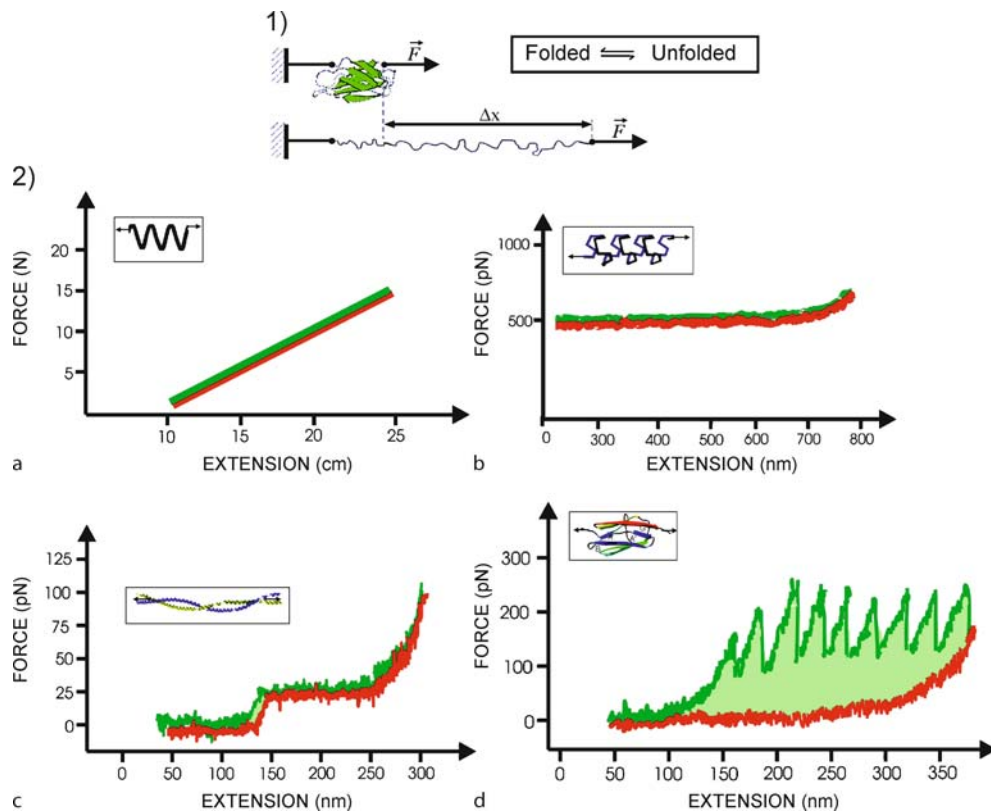
Single-molecule force spectroscopy (SMFS) by AFM: physical principle and modes of operation. **a** Schematic diagram of the **typical AFM** used in pulling experiments. The *line* depicts the laser light path before (*dashed line; pale red*) and after (*solid lines; red*) pulling on the system (i. e., protein, *green*). The protein connects mechanically the tip and the substrate, which is in turn bound to a piezoelectric positioner in this specific setup. The movement of the positioner along the z-axis results in bending of the cantilever along the same axis. This bending is tracked by changes in the reflected angle of a laser beam bounced off the cantilever, which in turn is detected by a split photodiode as a voltage difference between the two channels and is converted into force using Hooke's law. **b** 1 Typical force curve diagram in length-clamp SMFS mode showing different snapshots of the movement of the cantilever and tip as the positioner completes an **approach-retraction cycle**: it starts with the substrate not in contact with the tip (1), then it contacts with it, which bends the cantilever (2) increasingly (3); afterwards it is withdrawn from the tip, which bends the cantilever the other way (4) as it adheres to the tip, originating a force peak (4) on "jumping off contact" from the tip (5) which ends with the substrate again not in contact with the tip (6). 2 Schematic force-extension diagram showing a recording of a typical sawtooth pattern obtained by stretching of a multidomain protein molecule. **c** Force-clamp mode of SMFS showing the typical staircase extension-time recording. This particular example shows a polyubiquitin protein (N-C linked) being stretched at a constant force (110 pN). (Modified from [27], © Springer-Verlag 2006, with permission from Springer Science and Business Media)



“force spectroscopy” or “force-measuring” configuration of the AFM (Fig. 2) was designed to record force-extension curves obtained by pulling in a single direction ( $z$ -axis). Single molecules can be readily analyzed in this way by “single-molecule force spectroscopy” (SMFS).

SMFS is a very sensitive technique that can measure forces of tens of piconewtons and changes in length at nanometer resolution. A common problem is that force peaks can originate from a variety of sources other than the interaction of interest (detachment of other molecules from any of the two anchoring points, protein-protein in-

teractions, disentanglement of molecules, etc.), or from multiple molecules in parallel. This serious drawback was overcome in pioneering studies of protein nanomechanics by using modular proteins [87,94] or homomeric recombinant polypeptides [25], in which their pseudo-periodicity or pure periodicity, respectively, was used to infer single molecules unequivocally (i. e., a single-molecule reporter). These protein molecules are first immobilized between the substrate (a glass coverslip) and the force sensor so that a “mechanical circuit” is established that connects these two points (Fig. 1). Typically, in these experiments pro-



#### Protein Mechanics at the Single-Molecule Level, Figure 3

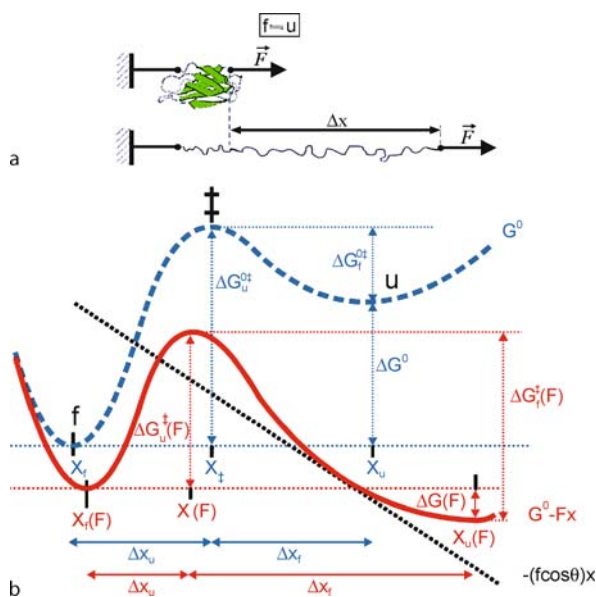
Elastic behavior of proteins under SMFS (length clamp mode). 1 Cartoon representation of the process of mechanical unfolding for a protein module using an applied force ( $F$ ) that results in an extension of the protein ( $\Delta x$ ). 2 Equilibrium and nonequilibrium in protein unfolding-refolding (green unfolding process, red retraction, pale green hysteresis; insets show the corresponding structures). a Force-extension curve of a “Hookean” spring such as the AFM cantilever (a macroscopic system). b Force-extension curve of an elastomeric protein: elastin. This protein behaves as an entropic spring in equilibrium, showing no hysteresis. c Force-extension curve of an elastic protein (inset): the tail of myosin II. This structure behaves close to a truly elastic protein showing little hysteresis when relaxed, which reflects the small amount of energy dissipated between extension and retraction. d Force-extension curve of a modular protein: a region of 8 Ig modules from the elastic region of titin. This structure shows a characteristic entropic-enthalpic force spectrum. Stretching the ends of a multimodular protein sequentially unfolds the domains, generating a typical saw-tooth pattern. The forced unfolding of this structure is a nonequilibrium process (note the marked hysteresis), which makes this protein a perfect shock absorber: it dissipates as heat part of the energy put into the system by the external mechanical work. Inserts: structure of a typical macroscopic spring (a), putative structure of elastin (b), structure of myosin II tail (c), and structure of the titin I27 Ig module (d). (Modified from [27])

teins become attached by physisorption (i. e., nonspecific adsorption via “physical forces”) to the two elements, although sometimes specific functionalization methods are also used (e. g., terminal cysteine residues in the protein to covalently link it to the gold coated surface of the substrate or/and the cantilever tip). Often, proteins in solution are attached nonspecifically to the substrate forming a dense layer of molecules from which the cantilever tip can pick out single molecules at random. Protein molecules are then stretched as in a lilliputian medieval rack, by moving apart the AFM positioner (bound to the substrate), which applies a mechanical force that unfolds the individual domains. This imposes a specific reaction coordinate (i. e. the end-to-end distance) on the unfolding process. By retracting the AFM positioner, the protein can also be refolded in the presence or absence of mechanical force. These experiments are typically performed under nonequilibrium conditions (Fig. 3). There are two basic modes currently being used depending on the variable being controlled: the most common, length-clamp, yielding a “force-extension” curve; and force-clamp, which yields an “extension-time” curve (Fig. 2).

### What Kind of Information Can Be Extracted from Stretching Proteins?

When the length-clamp mode is used on a modular protein or a polyprotein, the first source of resistance to extension typically comes from entropic forces, as these are “polymeric” proteins formed by several “pseudo-repeats” or perfect repeats, respectively. Entropic elasticity is a general property of polymers that results from the tendency of a chain to form a coil in order to maximize the conformational freedom (entropy) of its constituent monomers under the drive of thermal fluctuations. Stretching this polymeric molecule reduces its entropy producing a restoring force that bends the cantilever. In contrast to the linear relationship that governs cantilever bending (common mechanical springs follow Hooke’s law, Fig. 3), the entropic elasticity of a protein follows a non-linear relationship that can be formally described by the so-called “worm-like chain” model of polymer elasticity (WLC, see Sect. “Mechanical Dissection of Titin I27 Module: A Model System”).

Further extension of the protein may unravel it in a typically all-or-none fashion, which can be described by a two-state model (Fig. 4, see below). This exposes previously “force hidden” amino acid residues to force, resulting in an increase in the end-to-end length of the protein trapped between the tip and the positioner ( $\Delta L_c$ ), while relaxing the force acting on the cantilever to near



**Protein Mechanics at the Single-Molecule Level, Figure 4**

The effect of a mechanical force on the energy diagram of a protein. **a** Cartoon representation of the process of mechanical unfolding as in Fig. 3a. **b** The effect of a mechanical force on the free energy diagram of a protein that unfolds following a two-state model (f folded, u unfolded). The dashed blue curve represents the process in the absence of an applied force. An applied force (black dashed line) tilts the energy diagram of the process, decreasing the barrier to the transition state,  $\ddagger$  [ $\Delta G^\ddagger(F) < \Delta G_0^\ddagger$ ] (red line). The application of force also lowers the energy of the unfolded state relative to that of the folded state [ $\Delta G(F) < 0$ ]. The mechanical reaction coordinate is  $x$ . With an applied force, the positions of the free-energy minima ( $x_f$  and  $x_u$ ) and the maximum ( $x^\ddagger$ ) shift such that  $\Delta x_u$  becomes shorter ( $\Delta x_u, \text{red} < \Delta x_u, \text{blue}$ ) and  $\Delta x_f$  longer ( $\Delta x_f, \text{red} > \Delta x_f, \text{blue}$ ). Local curvature of the free-energy surface dictates these relative shifts in position. (Modified from [19] © Annual Reviews 2004, with permission)

zero (see Sect. “Mechanical Dissection of Titin I27 Module: A Model System”). Thus, unfolding events are detected as sudden changes in the end-to-end distance of the molecule. For a modular protein or a polyprotein, further extension repeats this cycle for each of the remaining folded modules, giving rise to a characteristic sawtooth pattern in the force-extension curve (Fig. 2b). When the force-clamp mode is used, the extension-time curve adopts a staircase pattern (Fig. 2c).

SMFS permits the direct measurement of both the mechanical stability of the barriers that a protein offers to its stretching (i. e., the mean force, as the process of unfolding is thermally driven and is stochastic) and the location of these barriers with single amino acid resolution [24]. As we shall see, the mechanical stability of a protein is a property that cannot be inferred from thermodynamic stability

measurements. Additionally, and based on conventional transition state theory, we can infer the kinetic parameters of the forced unfolding/refolding reaction [9,40]. An applied mechanical force tilts the energy diagram of the process, decreasing the barrier to the transition state and increasing the rate of the forward reaction (Fig. 4).

It is possible to calculate the probability density for unfolding, which predicts the most likely force of unfolding in terms of the spontaneous unfolding rate constant as follows:

$$F_u = (k_B T / \Delta x_u) \ln(r \Delta x_u / k_u^0 k_B T).$$

By using this analytical solution we can calculate the kinetic parameters for the unfolding reaction:  $k_u^0$  (spontaneous rate of unfolding) and  $\Delta x_u$  (width of the activation energy barrier, i. e. the distance on the reaction coordinate over which the force must be applied to reach the transition state). This equation predicts that the mechanical stability of a protein ( $F$ ) depends on the unfolding distance,  $\Delta x_u$ , the height of the barrier ( $\Delta G_u^\ddagger$ , which depends on  $k_u^0$ ), thermal energy, and the loading rate used during the extension of the protein ( $r = dF/dt = k \cdot v$ ; where  $v$  is the pulling speed).

Also, by relaxing the tethered protein before it breaks and waiting for appropriate periods of time, we can perform refolding experiments (in the presence of force or in its virtual absence) from which we can extract the equivalent parameters of the folding process.

Two recent SMFS studies show that it is now possible to gather both detailed information on protein structure, through a method called “mechanical triangulation” [38], and to probe dynamic rearrangements within the active site of an enzyme with unprecedented resolution [129].

### Computer Simulations of the Mechanical Unfolding of Proteins

SMFS experiments can measure changes in length at single amino acid resolution [24]. However, the typical methods used cannot provide a detailed structural resolution nor a “microscopic” (in the physicist sense, “nanoscopic” for biologists) interpretation of these processes. To this end computer simulations by molecular dynamics have proven to be very important for the atomic analysis of this process. They allow us to follow how mechanical forces change the structure of proteins under stress at the atomic level. These simulations are relatively simple to implement since the denaturing agent (force) and the reaction coordinate (N-C distance, if the protein is pulled by its termini) can be readily simulated. As these techniques also deal with single

molecules, they are especially suitable for direct comparison with SMFS results, which constitutes one of the main appeals of SMFS for theoreticians. The methods that have been used are all-atom (with and without explicit water solvent) and coarse-grained (e. g., on- and off-lattice simulations) models. The latter approaches give less detail but allow massive studies to be carried out at near experimental pulling speeds.

Classical all-atom simulations have been used for decades as an approach to interpret the behavior of proteins near their native states over time scales up to the order of 100 ns, with and without an explicit water solvent [46,50,93,109]. The usual approach is to resolve Newton’s equations (for instance, by using the leap-frog or prediction-correction algorithms) for all of the atoms in the system, monitor trajectories, and calculate time averages. However, interatomic forces have many sources. The forces are typically described by about 1000 parameters, which are collectively known as the ‘force field’ and they are determined through studies of model peptides. The primary contributions are Coulombic interactions between partial charges placed on the individual atoms. The all-atom models are often considered as being realistic representations of proteins, however, they incorporate several important approximations. Indeed, they neglect the dynamics of electrons as described by Schrödinger’s equation and what is often more important, the force field is ‘trained’ to be valid near expected native states of simple proteins and not in their non-native conformations.

By applying all-atom simulations to protein stretching we can obtain a wealth of information about the atomic details of conformation, which often enables the nature of mechanical stability to be elucidated (i. e., to explain which modules give rise to resistance to pulling). However, in order to fit the process to the computationally available time scales, very fast pulling speeds are used. These simulations are generally performed at pulling speeds many orders of magnitude faster than those of the experiment (eight or more, typically 1–10 mm/ms vs. 0.1–1 nm/ms). Because of this reduction in time, additional irreversible work is done to stretch the molecule, which typically results in force peaks one order of magnitude higher than those observed in AFM experiments. Hence, as simulations not fully reproduce the experimental conditions, it is not clear whether they precisely model the real process in these short periods of time in which new physical phenomena may arise. It should be noted that the peak heights at the computational speeds are usually substantially larger not only than the experimental unfolding force peaks but also than their extrapolation, assuming a logarithmic dependence on the speed that is often found.

These simulation techniques have a much higher time resolution than experimental single-molecule techniques (which can only capture slow conformational motions). Computational times are also shorter, in the range of nanoseconds or tenths of nanoseconds, which are achieved by increasing the pulling speed in the case of constant-velocity simulations (or by increasing the force in the case of constant-force simulations). However, a direct comparison is not possible because these techniques simulate the unfolding process over a very short time period (picosecond to nanosecond range), whereas experimental SMFS data are obtained over much longer times (millisecond-second range).

In spite of these difficulties, the synergy between all-atom simulations and experiments is very powerful. It is expected that such simulations will be important for the future development of the field in order to obtain the necessary high resolution picture of the mechanical unfolding of proteins. The first protein for which stretching was simulated was the titin I27 domain, which has become a model system in the field (see Sect. “[Mechanical Dissection of Titin I27 Module: A Model System](#)”). So far, about 30 protein structures have been studied using all-atom simulations, representing about half of the proteins folds studied by SMFS. Both numbers are only a tiny fraction of proteins stored in the PDB or existing in any organism. In this proteomic era there is a need for approaches to investigate the elastic landscape of large numbers of proteins and coarse-graining has recently offered such a tool.

Coarse-grained molecular dynamics models provide access to the experimental timescales of protein processes, in an approximate way, not only by reducing the number of degrees of freedom (e.g., by dealing only with the  $C^\alpha$  atoms) but also by introducing simple effective interactions that pertain to the larger scale level of description. The coarse-grained models are generally well suited for studies of large conformational changes, such as those occurring during folding, stretching and thermal melting. They offer further advantages when considering biomolecular complexes, such as the multiple linkages of proteins and ribosomes [116]. Their high-throughput character allows the comparison of the properties of a large number of proteins. A further simplification of the model may be implemented by making the conformational space discrete through constraining the locations of the  $C^\alpha$  atoms to certain sites of a lattice. This step also requires replacing Newton-equations based dynamics by Monte Carlo sampling [74].

Coarse-grained models of proteins have gained popularity in the last decade, although biochemists tend to dismiss them because of their reduced ‘realism’. Among the

coarse-grained models of proteins, the so-called Go-like systems are currently the most widely used [17,32,35,52,59,63,120,126], particularly since they are easy to implement and yet specific to a protein. The general idea is to formulate a model with effective couplings that is consistent with the experimentally established structure of the native state [1,114]. Clearly, there is no unique prescription for how to implement it. However, there is a host of simple choices that produce fairly good correlations with the experimental SMFS data. Go-like models have questionable features to study folding, but they are expected to be more reliable for studies of stretching where much of the relevant dynamics of the process takes place near the native state.

The construction of a  $C^\alpha$ -based Go-like model starts by representing each amino acid by a bead located at the  $C^\alpha$  site. Consecutive beads are tethered by harmonic potentials with minima at 3.8 Å. The next step is to introduce bead-bead interactions that would minimize the potential energy of the system in the experimentally established native structure. This is done by first determining which interactions should be operational in the native state and which should not. This results in the generation of the so-called contact map, which lists the native contacts (i.e. interactions). A sensible way to construct the contact map is to read in the all-atom native structure and represent each amino acid by a cluster of ‘grapes’. Each grape has a radius equal to the van der Waals radius of the corresponding atom multiplied by a factor to account for attraction. If such clusters overlap in the native state a native contact arises, which represents specific interactions such as hydrogen bonds, ionic bonds, disulfide bonds, and so on. Otherwise the contact is non-native and the corresponding interaction is made repulsive to prevent entanglement.

Once the contact map has been determined, one needs to decide on how to pick an effective potential that represents a contact whatever its chemical nature. Among the simple choices, the Lennard-Jones potential,  $V_{LJ} = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6]$ , appears to work particularly well. Here,  $r$  is the distance between interacting  $C^\alpha$  atoms,  $\sigma$  is the length parameter, which is determined contact-by-contact so that the minimum in the potential coincides with that in the experimental structure.  $\epsilon$  is the energy scale. In its simplest version, the energy scale is uniform and equal to 1.3–1.6 kcal/mol (so that the unit of force,  $\epsilon/\text{\AA}$ , is 67–110 pN). Disulfide bonds are covalent in nature so their effective coefficient  $\epsilon$  is an order of magnitude larger. The model can be improved by terms that represent local backbone stiffness. They favor the native values of the bond and dihedral angles. In order to mimic the thermostating effects of the solvent, one also introduces random

forces whose amplitude is proportional to temperature,  $T$ , (Langevin noise) and a damping force that produces overdamping. One can take  $k_B T/\varepsilon \sim 0.3$  to represent room temperature situations. The characteristic time scale ( $\tau$ ), in the resulting molecular dynamics simulations is of the order of ns instead of ps since it relates to the diffusion time across a distance of a typical  $\sigma$ , instead of an oscillatory time.

It should be noted that temperature exerts a profound effect on the force-displacement patterns since thermal fluctuations contribute to unraveling. The higher the  $T$ , the smaller the force peaks. At sufficiently high temperatures ( $k_B T/\varepsilon \sim 0.8$ ), in the so-called entropic limit, the force peaks due to the contact potentials disappear altogether. It is in this limit that the physics of an entropic chain applies.

The stretching simulations in all-atom models are usually implemented by making the molecular dynamics “steered”, which means that pulling is done directly through the attachment to a selected amino acid, usually a terminus, and this amino acid is made to move in a specific way. In coarse-grained models, stretching is usually accomplished by attaching two amino acids to elastic springs that imitate the elasticity of the cantilever on one end and that of the attachment to a substrate on another end. These simulations can be done at the experimental pulling speeds although, in order to survey the PDB at reasonable times, they are typically done two orders of magnitude faster.

### Protein Engineering is a Fine Tool for Protein Nanomechanics

Protein engineering has been critical for the progress of this field not only because it allows the construction of polyproteins (single-molecule markers for SMFS) but also for permitting the introduction of mutations.

Mutational analysis is a high-resolution tool that permits minimal perturbations of the system to be introduced in order to analyze the underlying molecular details. It has offered a unique way to study the molecular basis of how proteins respond to mechanical force. Several types of mutant proteins of the titin I27 domain (See Sect. “[Mechanical Dissection of Titin I27 Module: A Model System](#)”) have been used in SMFS studies. Loop insertion using glycine residues was used to show the existence of a “mechanical clamp” in I27 formed by backbone hydrogen bonds, demonstrating also the amino acid resolution of the technique [24]. Proline mutagenesis was used to show the existence of a mechanical folding intermediate [75] and to alter the mechanical stability of

the domain [69]. This approach takes advantage of the fact that proline (a known breaker of secondary structure) is an imino acid and as such, cannot form backbone hydrogen bonds. Conservative substitution and deletion mutagenesis have been used to demonstrate the existence of the I27 folding intermediate [45]. Furthermore, analysis of the so-called mechanical  $\Phi$ -value holds promise for a more detailed examination of the structure of the transition state in a forced unfolding reaction [11]. This is the mechanical equivalent to the well established  $\Phi$ -value analysis [43], which is used to examine the conformational effect of a mutation based on the relative changes in free energy of the native, transition and denatured states ( $\Delta\Delta G_{N-\ddagger}$ ,  $\Delta\Delta G_{U-\ddagger}$ ). The  $\Phi$ -value is obtained from the changes in thermodynamic stability measured from the shifts in the equilibrium denaturation curves. The unfolding  $\Phi$ -value, defined as  $\Phi_u = \Delta\Delta G_{N-\ddagger}/\Delta\Delta G_{N-D}$ , is determined by comparing the effect of the mutation on the transition state and it measures the amount of native structure that is present around the mutated residue in the transition state. The  $\Delta\Delta G_{N-D}$  is determined using equilibrium denaturation experiments. This analysis can only be applied when  $\Delta x_u$  remains the same for wild type and mutant proteins. Using these methods it has been shown that mechanical unfolding of titin I27 is a 3-state process in which the first transition state (between the native and the intermediate states) is very similar to the native state (Fig. 5).

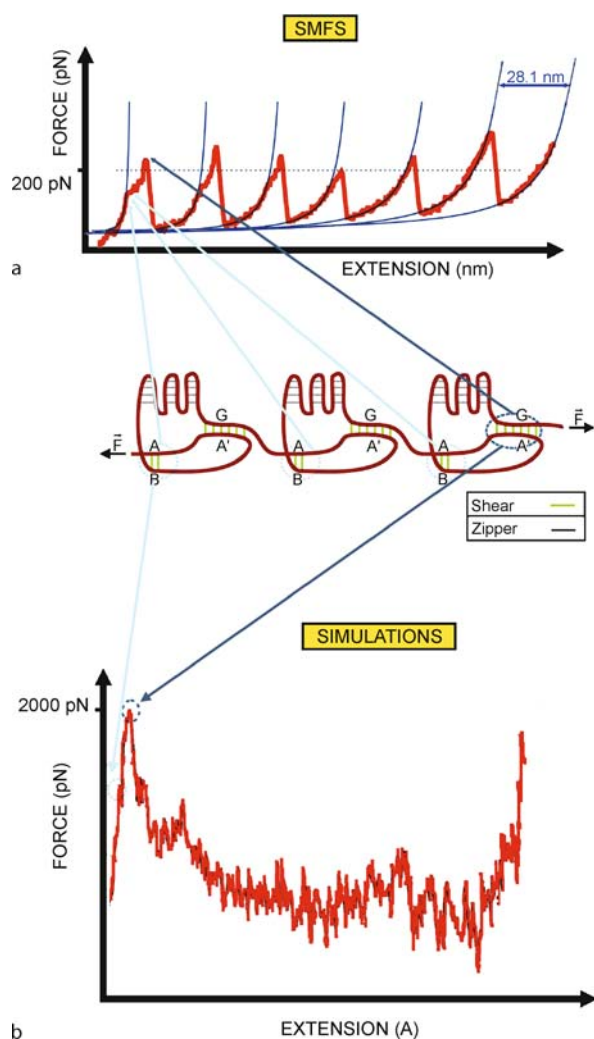
## Main Findings

### Which Proteins Have Been Studied?

Over the last decade both mechanical and non-mechanical proteins have been studied by mechanical unfolding and folding using SMFS [27,86] (Table 2). The protein structures analyzed so far include the following:

- the “all- $\beta$ ” protein E2Lip3 [17];
- several “all- $\beta$ ” structures of the  $\beta$ -sandwich type including “Ig-like” domains (from titin, fibronectin, tenascin, filamin, polycystin-1, Sls-kettin and projectin) and a “C2 domain-like” module (from synaptotagmin I) [18,26,70,87,89,92,105];
- a protein of the  $\beta$ -barrel type (GFP: green fluorescent protein) [37,90];
- several “ $\alpha + \beta$ ” proteins: ankyrin B, T4 lysozyme, barnase, DHFR (dihydrofolate reductase),  $\beta$ -grasp proteins (ubiquitin, protein L, GB1), RNase H, maltose binding protein, and Top7 [3,8,10,21,23,28,67,103,106,131];





- e) several “all  $\alpha$ ” structures (calmodulin, spectrin, and myosin II tail) [7,26,64,95,104]; and
- f) several unstructured proteins (elastin, titin PEVK, and titin N2B) [65,70,71,100,118].

Most proteins analyzed to date (i.e., modular proteins and recombinant polyproteins) have been pulled in the N-C direction, which is the natural direction provided by the peptide bond as synthesized at the ribosome. Exceptions are naturally occurring proteasomal polyubiquitins (K48-C linked) [23], E2Lip3 domain (K41-C) [17], lysozyme polyproteins cysteine-linked in solid state (C21-C124) [131] and GFP polyproteins cysteine-linked in solution (C3-C132, C3-C212, and C132-C212) [38]. These alternative linkages offer unique opportunities to apply force to different points of the protein. As we will see, these alternative geometries of pulling greatly alter the mechanical stability of proteins.

#### Protein Mechanics at the Single-Molecule Level, Figure 5

A model system in protein nanomechanics: the I27 module. Schematic representation of the mechanical architecture of the I27 module from titin. A three repeats polyprotein (center panel), showing patches of backbone hydrogen bonds in both “zipper” (dashed grey lines) and “shear” (AB, A’G; solid green lines) mechanical topologies. **a** A force extension-curve (force spectrum) obtained after stretching an I27 polyprotein by SMFS-AFM. **b** Force spectrum predicted by steered molecular dynamics (SMD) simulations, by an all-atom method. The main stability determinants of this module are a minor mechanical barrier (A-B patch of backbone hydrogen bonds in the polypeptide backbone) and a major one (A’G patch of backbone hydrogen bonds). The major resistance barrier corresponds to the force peaks in the force–extension spectrum (a). The minor barrier is detected in SMFS as a deviation of the WLC (“hump”), which is more apparent in the first peak of the spectrum since it collects the contribution of all the events from the stretched modules (pink arrows). This intermediate refolds after each unfolding event, as the force relaxes. These results are correlated with SMD simulations, where a low force peak and a high one can be observed when stretching a single domain (b). Although the extension data are similar (28.1 nm in AFM vs. 25–30 nm in SMD), the force data are different. Thus, the AFM mean peak value is about 200 pN, whereas the main force peak in the SMD simulation is about 2,000 pN. This discrepancy can be attributed to the different timescale of the experiments; simulations are much shorter processes (1 ns to tenths of 1 ns) than experiments (milliseconds to seconds) and thus they cannot fully reproduce the experimental conditions. (Modified from [27], © Springer-Verlag 2006, with permission from Springer Science and Business Media)

In addition to single polypeptides, the mechanical properties of some protein complexes as a whole have been also analyzed by SMFS (e.g., myosin II tails [104] and *E. coli* adhesive pili [81]).



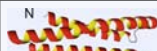

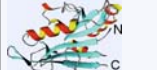

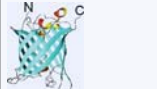
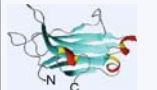

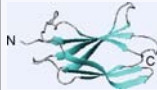

#### Molecular Determinants of the Mechanical Stability of Proteins

The number of proteins analyzed so far by SMFS is still very small (about 23 different proteins, roughly 55 PDB structures) and they have been studied with differing detail (systematization of the conditions in which these studies are done would be highly desirable, as already proposed [97]). Although this prevents us from being able to extract general principles, regarding mechanical stability some tendencies have been observed that can be summarized as follows [27,86]:

Proteins have widely different mechanical stabilities, measured by the most probable unfolding force ( $F_u$ ) when pulled in the N-C direction: they range from less than the limit of resolution of AFM (typically  $\sim 20$  pN; e.g. calmodulin) to 330 pN (e.g., titin Ig domains). Interestingly, mechanical proteins intended to resist force tend to be more mechanically stable than non-mechanical ones or the so-

**Protein Mechanics at the Single-Molecule Level, Table 2**

Mechanical stability of representative proteins. Mechanical stability (measured in pN) is defined as the average force of unfolding. The proteins listed here were all pulled apart from their N- and C-termini at comparable loading rates

Protein fold		Examples	Mechanical stability, F (pN)
<b><math>\alpha</math>-helix</b>		calmodulin	< 25
<b><math>\beta</math>-spiral</b>		$\alpha$ -elastin, PEVK (titin)	< 25
<b><math>\alpha</math>-helical structures</b>			
bundles		spectrin, dystrophin, myosin tail, ankyrin B repeats	25–50
solenoids		ankyrin B	360 ?
<b><math>\alpha + \beta</math> structures</b>		dihydrofolate reductase, barnase	27–70
<b><math>\beta</math>-grasp</b>		ubiquitin, GB-1, protein L	136–200
<b><math>\beta</math>-barrel</b>		GFP	100
<b><math>\beta</math>-sandwich</b>			
“zipper”		C2A domain	60
“shear” (PKD)		polycystin-1	50–250
“shear” (fn)		titin A band, fibronectin, tenascin, projectin (fn), myomesin (fn)	80–200
“shear” (Ig)		titin I band, sls-kettin, projectin (Ig), myomesin (Ig), filamin A	150–330

called elastomeric proteins, although there are some exceptions.

Unstructured and  $\beta$ -spiral proteins (e.g. elastomeric proteins like elastin and the PEVK and N2B regions from human titin) are among the less mechanically stable proteins. Such proteins are entropic springs and behave reversibly.  $\alpha$ -helical proteins (e.g., calmodulin, T4 lysozyme) also have relatively low mechanical stability although  $\alpha$ -helical bundles (e.g., spectrin, myosin II tail) and solenoids (e.g., ankyrin B) are more stable.  $\beta$ -stranded proteins tend to unfold at higher unfolding forces than  $\alpha$ -helical ones.

The mechanical stability of most mechanical proteins (e.g., Ig-like  $\beta$ -sandwich and the ubiquitin-like  $\beta$ -grasp folds) tends to be determined by highly-localized mechanical clamps at the breakpoint formed by shear hydrogen bonds between the backbones of the  $\beta$ -strands. However, in some protein folds (e.g., fibronectin type III, fnIII, from human tenascin- TNfnIII3) the hydrophobic core also contributes to mechanical resistance [85]. Exceptions to the shear mechanical clamp topology are GFP, ankyrin B, and the de novo designed Top7 fold. In addition to secondary-structure based elasticity, there are two more types of structure elasticity: tertiary (e.g., the solenoid of ankyrin

B [67]) and quaternary (e. g. the helical rod of *E. coli* adhesive pili [81] and the myosin II tail [104]).

Since force is a vector quantity, the mechanical stability and the mechanical unfolding pathway depend on the pulling geometry, which is affected by both the topology at the breakpoint and the point of application of the force. Hence,  $\beta$ -stranded proteins with a shear mechanical topology at the breakpoint (i. e. the force vector is orthogonal to the hydrogen bonds) are more mechanically stable than zipper  $\beta$ -stranded proteins (where the force vector is parallel to the hydrogen bonds). The points of application of the force to a protein are also relevant as they can substantially alter its mechanical stability [17,23], implying that proteins have “Achilles’ heels” in their structure. For instance, GFP pulled from geometries other than N-C shows mechanical stabilities of up to  $\sim 550$  pN [39].

Mechanical stability is a kinetic property which in general is not correlated with thermodynamic stability ( $\Delta G$ ) or with melting temperature ( $T_m = \Delta G/\Delta S$ ) [26]. Mechanical stability seems to be roughly predicted by the unloaded unfolding rate constant [27].

Chemical and mechanical unfolding have been shown to follow different pathways [10,45] and have different unfolding barriers [10,16].

Moreover, it has been demonstrated that mechanical stability can be modulated by ligand binding [3,58] and by disulfide bond formation [4,14,22,128,129].

Taking together, these findings show that proteins display a broad range of responses to mechanical stress, which cannot easily be rationalized in terms of predictors of mechanical resistance. Thus, although the molecular basis underlying the mechanical resistance of proteins is still unclear, several determinants have been identified through these studies: amino acid sequence, mechanical topology, unloaded unfolding rate constant and pulling geometry.

The molecular structure of a protein, poses constraints on the location of the transition state in mechanical unfolding. It has been suggested that tertiary interactions have shorter distances to their transition states than secondary structures, and they tend to be more brittle (i. e. they break at high forces and after small deformations) than secondary interactions, which are more compliant (breaking at low forces and after large deformations). Furthermore, tertiary interactions may require more time to equilibrate than secondary ones and therefore, they often present hysteresis in the pulling-relaxation cycle [19]. Most proteins show a high degree of connectivity and as a result, their unfolding seems to be highly cooperative with the stability of secondary structures depending on their tertiary context and often presenting no intermediates. Still, due to the local action of the applied force, their

mechanical stability tends to be related to highly localized molecular structures near the mechanical “breakpoint” rather than to the global structure [19,73]. A massive survey has been carried out recently to identify these mechanical clamps in all protein modules (up to 150 amino acids long) for which there are atomic structures available [110,111].

### Mechanical Dissection of Titin I27 Module: A Model System

The model system most commonly used to study mechanical unfolding/refolding in proteins is the I27 module from titin (Fig. 5), a gigantic multimodular protein responsible for the so-called passive elasticity of muscle (see Sect. “The Elasticity of Muscle Explained at the Single-Molecule Level” [117]). This module has an 89 amino acids long Ig-like  $\beta$ -sandwich fold (Fig. 3d and Table 2). All-atom molecular dynamics simulations of its stretching identified two patches of backbone hydrogen bonds as true “structural” barriers with different mechanical resistances: a low force barrier involving 2 hydrogen bonds between  $\beta$ -strands A and B (AB patch), and a high force barrier between  $\beta$ -strands A' and G involving 6 hydrogen bonds (A'G patch) [73]. The hydrogen bonds in both patches are perpendicular to the direction of the force vector (a “shear” mechanical topology: bonds are arranged “in parallel” in the mechanical circuit), whereas the remaining hydrogen bonds in the structure are parallel to the force vector (a “zipper” mechanical topology where the bonds are arranged in series) and like the hydrophobic core, seem to offer little resistance to extension (Fig. 5). Interestingly, coarse-grained models yield similar results [32,126].

These predictions were remarkably consistent with the experimental data obtained from SMFS using polyproteins: an intermediate found at low force ( $\sim 100$  pN, at 0.3–05 nm/ms) was associated to the rupture of the AB patch [75], whereas a high force peak ( $\sim 200$  pN, at 0.6 nm/ms) was found to depend exclusively on the A'G patch [25] and the associated side-chain packing (i. e. hydrophobic) interactions between the A' and G strands. The hydrophobic core of this structure plays no role in resisting force [13]. The hypothetical role of the A'G patch as a mechanical clamp was tested by loop insertion [24] and proline mutagenesis [69], these mutants providing the first mechanical phenotypes.

### Unveiling Intermediates and Rare Misfolding Events

As mentioned in the last section, experimental stretching of I27 also confirmed the existence of a weaker mechanical barrier in the AB patch, which appeared as a small

deviation in the force-extension recordings (at  $\sim 100$  pN and at 0.3–0.5 nm/ms) from the pure entropic behavior described by the WLC model (a “hump” of decreasing intensity on each of the saw teeth; Fig. 5). This deviation was interpreted as evidence for the existence of an unfolding intermediate involving the rupture of the AB patch. The existence of this unfolding intermediate in the wild type I27 module was later confirmed by using mutant proteins [45,75].

During refolding experiments using polypeptides or modular proteins, “skip” events are occasionally observed, whose contour lengths suggest they are due to the formation of a “superfold” which includes two consecutive domains plus the linker region between them. This misfolded “superdomain” unfolds at similar forces to those for a single domain and refolds back to the dimensions and mechanical stability of two normal domains. SMFS of the titin poly-I27, the fnIII domains of tenascin, and the R1617 tandem repeat of spectrin have demonstrated that these single-molecule studies can detect rare events in as little as 2, 4 and 3% of the population, respectively. Hence, the efficiency (i. e. fidelity) of refolding after mechanical unfolding can be estimated [21,88], which has been found to be exceptionally high for the non-mechanical protein GB1, a structural homologue of protein L (above 99.8%) [21]. Such rare unfolding events could not be detected using ensemble techniques and they open the door to investigate the means of reversing the undesirable misfolding that occurs in a number of pathological conditions.

### The Elasticity of Muscle Explained at the Single-Molecule Level

The basic contractile unit of muscle is the sarcomere. This relatively simple contraction/extension machine (i. e. it works in a single dimension) is highly elastic [55]. The pioneering work by Wang et al. [124] and Maruyama et al. [76] demonstrated that the so-called “passive elasticity” of muscle (i. e. generation of restoring forces that resist stretch independently of ATP) is mainly mediated by titin, a giant protein ( $> 3$  MDa, the longest polypeptide known to date) that spans half a sarcomere ( $\sim 1$   $\mu$ m; from the Z disk to the M line) and that acts as a molecular spring (Fig. 6). Passive elasticity plays an important role in muscle function since, typically, a muscle actively contracts against the elastic strain of a passively elongating muscle. This property ensures that the sarcomere recovers its initial dimensions on muscle relaxation [117].

A remarkable feat achieved using SMFS was the reconstruction of the passive elasticity of intact myofibrils by simply scaling up from the mechanical properties of sin-

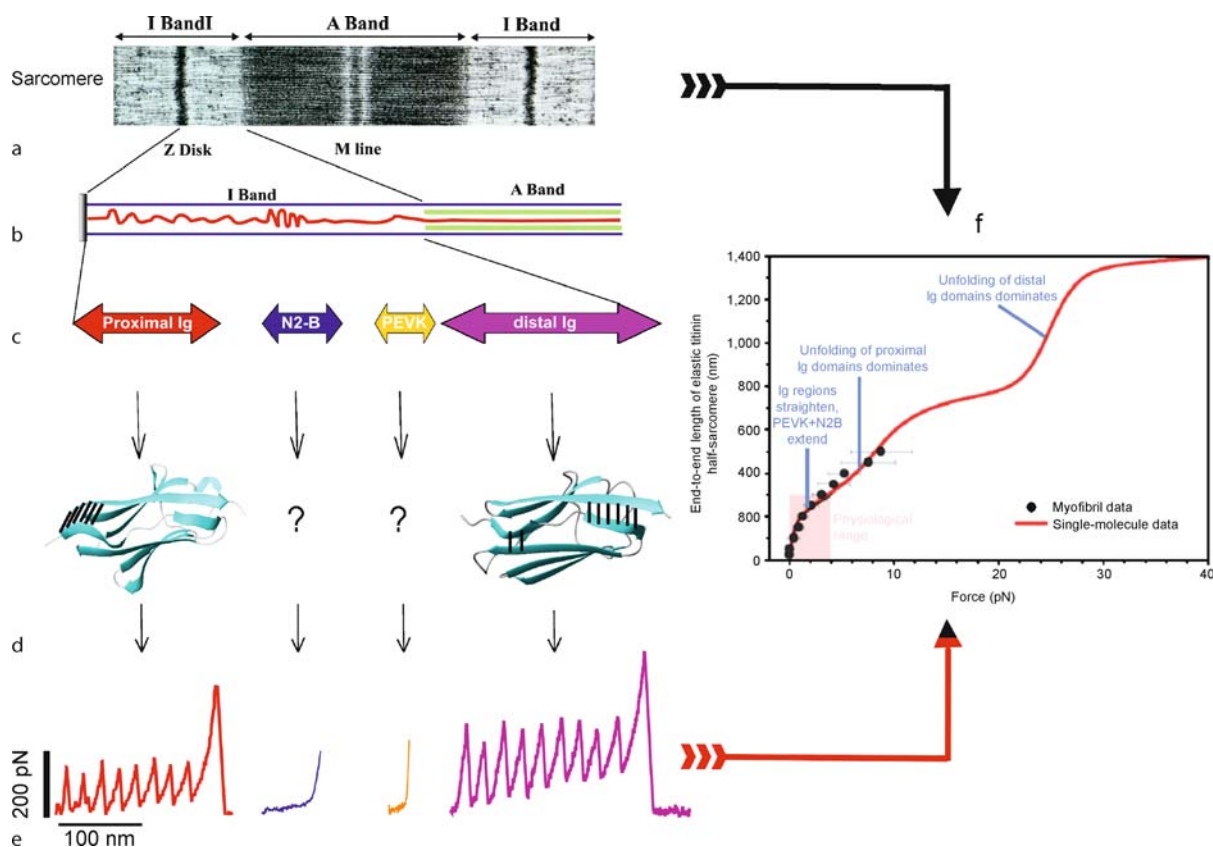
gle titin molecules [70] (Fig. 6). These were reconstructed from the mechanical properties of representative elements of its elastic region (i. e. N2B, PEVK, and the proximal and distal Ig regions relative to the N-terminus of the protein). These results showed that titin behaves very differently from a Hookean spring. Instead, in response to axial tension, titin behaves as a multistage non-linear spring that adjusts both its length and apparent stiffness by virtue of its particular modular design. At low forces the entropic springs dominate (i. e. N2B, PEVK, and Ig straightening) while at high force (i. e. in non-physiological conditions) the enthalpic springs (partial and total unfolding of a few Ig domains) act as “shock absorbers” to prevent damage of the sarcomere. Thus, titin Ig shock absorbers are essentially a “safety mechanism” for the sarcomere.

Through this reductionist approach, it has been possible to explain muscle passive elasticity (a macroscopic biological property) from the additive mechanical properties of a single sarcomeric protein at the single-molecule level. The example of titin elasticity shows how single-molecule experiments can be used to elucidate, at a more fundamental level, the physiological function of a protein through a pure biophysical dissection of a complex hierarchical system.

### Molecular Shock Absorbers Galore

Mechanical proteins tend to be modular, and are often composed of assortments of modules of the same type (e. g. immunoglobulin, fibronectin), which frequently display distinct mechanical stabilities. The shock absorber effect of the Ig domains from titin has also been found in cell adhesion proteins like tenascin (in its fnIII modules), where it was proposed to extend the range and lifetime of cell-cell interactions [87]. In the case of fibronectin, another extracellular matrix protein, most of its (Ig-like) fnIII domains remain folded during the stretching of its matrix [2], similar to titin.

Protein domains from muscle (titin, myomesin, projectin and Sls-kettin), cell adhesion (tenascin, fibronectin, polycystin-1), cytoskeletal (filamin), and surface receptor proteins tend to belong to the Ig-like  $\beta$ -sandwich family of proteins and probably have somewhat similar mechanical topologies at the breakpoint. This superfamily of Ig folds, which includes Ig (titin, projectin, Sls-kettin), fnIII (tenascin, fibronectin, titin, projectin), E-set (filamin), and PKD (polycystin-1) types, consist of 7 stranded  $\beta$ -sandwich structures in which the N- and C-terminal strands are parallel to each other, pointing in completely different directions (i. e.,  $180^\circ$ , Table 2, Fig. 3). These modules seem to have evolved to withstand forces, when con-



**Protein Mechanics at the Single-Molecule Level, Figure 6**

Reverse engineering of titin. **a–f** Reconstruction of muscle passive elasticity from the mechanical properties of single titin proteins. The mechanical properties of representative elements of its elastic region, I band (**b**). **c** These elements are the proximal Ig domains (the crystal structure of I1 is shown in **d**, left panel), the N2B and PEVK (hypothetical random coils) and the distal Ig domains (the NMR structure of I27 is shown in **d**, right panel). **e** Representative force-extension recordings obtained after stretching I4–I11 (in red), an N2B (in blue), a PEVK (yellow) and I27<sub>8</sub> proteins (violet). **f** According to this model, within the physiological range of sarcomere extension (i. e. forces below 4 pN) unfolding would rarely happen and most of the elasticity of titin would result from the entropic elasticity of straightening the Ig domains in the I band, and of extending the random coils (PEVK and N2B). (Modified from [70,86])

nected in series, thanks to a shear mechanical topology of the hydrogen bonds at the breakpoint. This arrangement may provide these domains with considerable resistance to mechanical stretching. Thus, the Ig-like  $\beta$ -sandwich fold seems to be a platform that can tolerate higher mechanical stress than other folds. Within this class, the Ig (forces ranging from 150–330 pN, at 0.6 nm/ms) and the PKD (~200 pN; at 1 nm/ms) domains appear to be more mechanically stable than the fibronectin type III (fnIII; 75–220 pN, at 0.6 nm/ms) or the E-set (to which filamin belongs; 50–220 pN, at 0.37 nm/ms) domains [27].

As discussed, SMFS is often used for the mechanical analysis of multi-modular proteins and polyproteins. In these proteins, stretching results in the unwinding of many segments. Often the patterns are interpreted as a serial process. This seems justified in many cases (especially

for polyI27 or polyubiquitin), but there are examples when this assumption fails (e. g., polycalmodulin). Temperature also seems to play an important role in these experiments. At high temperatures, thermal fluctuations dominate and all segments unwind in parallel (i. e. simultaneously, because such fluctuations are independent of the segment number although they are sensitive to the sequential distance between amino acids that are in contact so that the short range contacts unravel last). Thus, in these conditions the weakest spots will unravel first, irrespective of whether they are hidden, exposed, or near the termini. On the other hand, at very low temperatures strong spots resist simultaneous unwinding, resulting in a serial process except at the very initial stages when all segments yield somewhat. At room temperature, there is a competition between serial and parallel unwinding pathways and either



one may win, depending on the nature of the modules. Thus, in general, one should always expect some degree of mutual influence between modules [29,31,56].

### On the Biological Meaning of Protein “Robustness”: Selected Trait or Epiphenomenon?

How well do SMFS experiments mimic protein mechanics *in vivo*? In the case of extension machines like the sarcomere, SMFS seems to adequately mimic the natural linear pulling geometry, since proteins are pulled apart from both ends of the polypeptide chain. This may also be the case for other cytoskeletal machines (the case of spectrin may not be as clear as it forms a mesh-like structure), the adhesion machinery, and some mechanosensitive ion channels. Furthermore, it has been suggested that some chaperonins may pull their protein substrates apart in a similar way prior to their refolding [107].

However, the case of the unfoldases is not so clear-cut (Fig. 1). The accepted model for protein translocases (from the mitochondrion, chloroplast and endoplasmic reticulum) and compartmental proteases (such as the proteasome), is also a mechanical one. Nevertheless, rather than a linear pulling geometry with two attachment points, the evidence here favors a different geometry that involves a single attachment point from which the pulling would be done by threading the protein towards the entrance of a narrow channel present in these nanomachines (i. e., in similar way to wire drawing in metallurgy). This model is mainly based on the fact that the susceptibility of substrate proteins to be unfolded by these nanomachines *in vivo* correlates more closely with the mechanical stability obtained by mechanical unfolding (using SMFS) than with thermodynamic or kinetic stability (measured *in vitro* by bulk chemical or heat denaturation). In the case of compartmental proteases, the AAA+ ATPase motor involved in the pulling process seems to unfold the structure adjacent to the degradation tag by trapping local unfolding fluctuations. Global unfolding then occurs immediately, driven by the cooperativity of the protein unfolding process [77,91,102]. As seen from SMFS findings, in this “local stability” model the structure and pulling geometry at the attachment point (i. e. local stability) are more important than their global counterparts. Similarly, protein import by the mitochondrial translocase depends on the N-terminal targeting sequence and the local structure of the adjacent protein, more akin to the vectorial nature of AFM pulling experiments than to solution or heat denaturation experiments [130]. Indeed, mechanical hypomorphic mutations can also accelerate mitochondrial import under specific conditions [101]. The vectorial nature of protein

translocation highlights the importance of the existence of Achilles’ heels in proteins. Accordingly, in order to unfold their protein substrates more economically, mechanical “unfoldases” might have evolved specific pulling mechanisms to take advantage of the presence of weak spots in the structure of the latter.

If the mechanical model is confirmed for “unfoldases” (molecular chaperones, compartmental proteases such as proteasomes, and the protein translocases from the protein import machinery of the mitochondria, chloroplasts, and endoplasmic reticulum), then most proteins in the cell (i. e., “mechanical substrates” of unfoldases) would be mechanically unfolded at some point or another during their lifespan, which would make their mechanical properties physiologically relevant.

Mechanochemical enzymes may not be alone in their capacity able to generate mechanical forces. Thus, according to the hypothesis of the tension-induced catalysis proposed by Haldane and Pauling, mechanical forces may also underlie the activity of other enzymes [19]. This hypothesis postulates that enzyme catalysis may work by inducing mechanical tension in the enzyme-substrate complex.

The correlation between mechanical stability and mechanical function holds well at high forces but it is less true below  $\sim 100$  pN [27]. This may reflect the possibility that for some proteins with no known mechanical function (e. g. protein L, GB1, GFP) their relatively high mechanical stability (in some cases derived from a shear mechanical topology at the breakpoint) may just be an epiphenomenon, unrelated to the biological function of the protein. Indeed, a number of the proteins studied may never experience mechanical forces in the cell and if they do, the pulling geometry *in situ* might be different to that of SFMS experiments. Alternatively, mechanical stability may simply be a neutral trait for certain proteins, not directly selected by evolution (i. e. a remnant by-product of their evolutionary history). For instance, the mechanical stability of fnIII domains from the titin A-band (non-elastic region bound to the myosin thick filament of the sarcomere and therefore probably not subject to axial stress) is only slightly lower than those of the I-band (180 vs. 200 pN, at similar pulling speed). Hence, if our current model of how titin works is correct, this trait does not appear to have been evolutionarily selected in the case of the A-band domains (non-elastic) but rather, it may be an epiphenomenon originated through constraints imposed during the evolution of this elastic protein. Alternatively, it is also possible that the anchoring geometry of these domains to myosin may require such a high mechanical stability.

In conclusion, by using SMFS only a certain type of mechanical stability in proteins can be accessed at present.

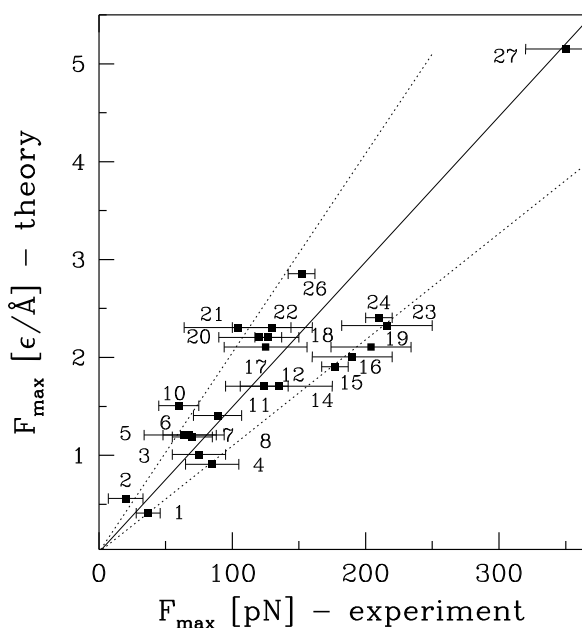
This mechanical stability (defined as  $F_{\max}$  when they are pulled from the N- to C-termini, using the length-clamp mode) would be expected to be an evolutionarily selected trait for components and substrates of mechanical biomachines that work like extension machines (i. e., similar to a medieval rack with two attachment points, Fig. 1a,b,c,d). However for other mechanical biomachines (e. g., wire drawing-like machines, Figure 1e) and for “non-mechanical” proteins, this parameter may well be a mere epiphenomenon. Thus, in each case careful comparative and functional analysis should be done to distinguish between these two possibilities.

### Predicting the Robustness of Proteins from Their Structure

As we have already pointed out, the nanomechanics of only a few proteins and protein modules has been studied to date. Considering how vast proteomes are, there are therefore plenty of mechanical proteins and unfoldase substrates waiting to be mechanically characterized. In the mean time, simulations provide a nice short cut in order to help us to predict and understand protein robustness, and to access pulling geometries that are not yet experimentally accessible (e. g., “wire drawing” pulling, [54,115,127]).

As mentioned above, using a simple variant of a Go-model a massive survey [110,111] has been recently conducted to identify mechanical clamps in 7510 protein modules, of up to 150 amino acids long, for which uncorrupted atomic structures were available at the time when the study was undertaken. Additionally, 239 substantially longer proteins were also included in the survey. This study has identified new mechanical clamps and has enabled the available protein structures to be ordered on the basis of their predicted mechanical stability. This model correlates fairly well (correlation coefficient of 89%) with the available experimental values of the mechanical stability,  $F_{\max}$ . The theoretical and experimental data points are listed in Table 3 and displayed in Fig. 7. The comparison is restricted to 28 experimentally stretched proteins for which the PDB structure is available.

This survey used a simple Go-like model and stretching was implemented at a constant speed and at a temperature meant to correspond to room temperature. The values of resistance to pulling,  $F_{\max}$ , together with the force-displacement patterns are available at [info.ifpan.edu.pl/BSDB](http://info.ifpan.edu.pl/BSDB) (Bio-molecule Stretching DataBase). The survey is restricted to stretching at the termini (i. e., N-t and C-t amino acids), although it should be kept in mind that there



**Protein Mechanics at the Single-Molecule Level, Figure 7**

Modeling mechanical stability with coarse-grained molecular dynamics. Correlation between the experimental and theoretical values of  $F_{\max}$  for a simple Go-like model with uniform Lennard-Jones potentials in the native contacts. The contact map is obtained by studying heavy atom overlaps and by eliminating the  $i, i + 2$  contacts which usually are weak. The chain of amino acids is also endowed with local terms that mimic the backbone stiffness. The numbers in the top left panel indicate particular proteins. These are: 1(1n11), 2(1cfc), 3(1hci), 4(<sup>10</sup>FNfnIII), 5(1u4q), 6(1aj3), 7(B), 8(1ubq(48-N)), 9(1b6i), 10(1rsy), 11(<sup>13</sup>FNfnIII), 12(<sup>12</sup>FNfnIII), 14(<sup>3</sup>TNfnIII), 15(1qjo(N-41)), 16(G), 17(<sup>1</sup>FNfnIII), 18(I1), 19(I27), 20(1emb), 21(1emb(132-212)), 22(1emb(3-212)), 23(1ubq), 24(1nct), 25(1g1c), 26(L), 27(1emb(3-132)), 28(1vsc). B, L, and G denote barnase, protein L, and protein G respectively. The solid line corresponds to  $\epsilon/\text{\AA} = 67$  pN and the lower dotted line to 111 pN. This figure also illustrates the best pick for  $\epsilon$  (1.3 kcal/mol, the unit of force 67 pN) and this translation of the theoretical results into values in pN is shown in Table 3. Adapted from [113]

is a significant linkage dependence, i. e.  $F_{\max}$  depends on the particular pulling geometry.

The first observation of this study is that the distribution of  $F_{\max}$  across the PDB is non-Gaussian and has a pronounced tail at the high end (the end corresponding to “strong” proteins). The I27 domain of titin is a fairly strong mechanical protein, with its  $F_{\max}$  being about twice as high as the average. However, there are a substantial number of proteins that are predicted to have twice the strength of this module. In 80% of cases, the maximum force is found at the early stages of stretching. The force was found not to depend on the number of amino acids ( $N$ ) as one can find weak and strong proteins for any value

**Protein Mechanics at the Single-Molecule Level, Table 3**

Predicted and measured mechanical stability in proteins. Comparison between experimentally measured values  $F_{\max}$  with theoretical predictions in the Lennard–Jones Go-like model. The theoretical results are averaged over 10 trajectories to account for several pathways if any. Proteins were pulled by their termini except for the ones in which the amino acids being pulled are indicated in brackets. The first column shows the PDB code and the second one the number of amino acids. Adapted from [113]

PDB	N	$F_{\max}$ [pN] -experiment	$F_{\max}$ [pN] -theory	
1tit	89	204 ± 30	144	I27*8
1nct	98	210 ± 10	161 ± 13	I54-I59
1g1c	97	127 ± 10	154 ± 13	I1 titin
1b6i	164	64 ± 30	80	T4 lysozyme(21-141)
1aj3	106	68 ± 20	82	spectrin R16
1qjo	80	15 ± 10	80	eE2lip3(N-C)
1qjo	40	177 ± 10	134	E2lip3(N-41)
1dqv	127	60 ± 15	100	calcium binding C2A
1rsy	127	60 ± 15	100 ± 13	calcium binding C2A
1byn	127	60 ± 15	94	calcium binding C2A
1cfc	148	< 20 pN	55 ± 20	calmodulin
1n11	33	37 ± 9	27	ankyrin*1
1bni	108	70 ± 15	94, 114	barnase/I27
1bnr	108	70 ± 15	70	barnase/I27
1bny	108	70 ± 15	74, 87	barnase/I27
1hz6	67	152 ± 10	235	protein L
1hz5	67	152 ± 10	188	protein L
2ptl	67	152 ± 10	147 ± 13	protein L
1ksr	100	45 ± 20	134 ± 20	DdFLN -4
2rn2	155	19 ± 10	121 ± 13	ribonuclease H
1ubq	76	230 ± 34	155	ubiquitin
1ubq	76	203 ± 35	155	ubiquitin(N-C)*9
1ubq	28	85 ± 20	60	ubiquitin(K48-C)*(2-7)
1emb	129	350 ± 30	345 ± 25	GFP(3-132)
1emb	219	130 ± 30	154, 288	GFP(3-212)
1emb	80	120 ± 30	147 ± 13	GFP(132-212)
1emb	235	104 ± 40	154 ± 13	GFP(N-C)
1fnf	94	75 ± 20	107, 121	<sup>10</sup> FNfnIII
1ttf	94	75 ± 20	47, 80	<sup>10</sup> FNfnIII
1ttg	94	75 ± 20	47, 67	<sup>10</sup> FNfnIII
1fnh	92	124 ± 18	121	<sup>12</sup> FNfnIII
1fnh	89	89 ± 18	94, 114	<sup>13</sup> FNfnIII
1oww	93	125 ± 31	141 ± 13	<sup>1</sup> FNfnIII
1ten	90	135 ± 40	114	<sup>3</sup> TNfnIII
1pga	56	190 ± 20	161 ± 13	protein G
1gb1	56	190 ± 20	111 ± 13	protein G

of  $N$ . However, the larger the  $N$  the higher the probability that force was higher. The  $\alpha$ -class of proteins usually showed weak forces. There are other specific correlations between types of structure, as described by the CATH classification (Class, Architecture, Topology, Homology), so that the strongest proteins correspond to two types of specific architectures: 30%  $\alpha$ ,  $\beta$ -rolls (ubiquitin-like) and 60%  $\beta$ -sandwiches (titin-like).

The unraveling process can be represented by “scenario diagrams” that show for how long a given native contact holds. Accumulation of such events results in the appearance of a force peak. Thus, it is possible to determine which contacts give rise to a force peak and then assess their dynamic impact by removing them in small sets and determining what effect this has on  $F_{\max}$ . The contacts whose removal reduces  $F_{\max}$  substantially correspond to

a mechanical clamp (i. e., a resistance point). In the top strongest proteins, 95% have mechanical clamps corresponding to long stretches of parallel  $\beta$ -strands that shear on pulling (like titin modules). Particularly large forces arise when such parallel strands are further stabilized by neighboring parts of the structure. However, there is also another type of mechanical clamps, those formed by antiparallel  $\beta$ -strands that are unstructured and even delocalized (where the clamping action sits in disjoint places).

Finally, there is a small but interesting group of proteins (about 300 examples so far) which contain simple knots in their native structures, usually trefoil knots. Knots are attractive topological features of proteins biologically relevant (e. g., toxicity), which are predicted to have important mechanical properties. These proteins could be pulled by a non-terminal amino acid to unravel the knot whereas pulling at the termini would tighten the knot. The latter case has been studied theoretically using the Go-like model for 20 proteins [112]. The simulations suggest that the knot tightening in a stretched protein proceeds through jumps, i. e. sudden displacements of the ends of the knot along the sequence. (The ends of a knot can be identified by removing the  $C^\alpha$  atoms as long as the backbone does not intersect a triangle set by the atom under consideration and its two immediate sequential neighbors.) These jumps have definite lengths and together with the final location of a tightened knot they are specified by the local geometry of a protein chain (sharp turns are favored). The larger the size of a knot the larger the number of jumps observed before its final tightening. However, such jumps are not observed in the dynamics of knot motion on stretched polymers. In this case, the motion has a diffusive character and usually results in sliding of the knot off the chain. Another possible way to manipulate such a protein is to pull it to a certain extension and then release it abruptly. If the stretching stage lasts sufficiently long (so that several force peaks are observed and the knot gets tightened substantially) then the protein misfolds upon release and the knot ends up residing at metastable locations (this is predicted to happen for 2etl ( $N = 223$ ), 1vho ( $N = 157$ ), and 1v2x ( $N = 191$ ) and in 80% of trajectories for 1o6d ( $N = 147$ ). However, the knot in protein 1j85 ( $N = 156$ ) was usually found (for most trajectories) to return reversibly to its native location.

Go-models can also be used to simulate the pulling of membrane proteins out of membranes (an important topic that is out of the scope of this review as it involves not only unfolding but also unbinding from the membrane components), resulting in a multipeak force pattern for bacteriorhodopsin that was remarkably similar to the experimental one [33]. All these results eagerly encourage the

use of Go-like models for comparative studies in a variety of proteins.

### Insights into Protein Folding from Forced Unfolding/Folding

Protein nanomechanics is multifaceted. One attractive element of our new capability to unfold/refold proteins by force is that it may provide new insights into the protein folding problem for any protein (i. e., mechanical proteins, “mechanical protein substrates”, and non-mechanical proteins).

Ten years ago the only way to measure the stability of a protein was to change its physical (temperature or pressure) or chemical environment (using guanidinium chloride or urea, or varying the pH), and to monitor the loss of protein conformation by spectroscopic techniques in order to determine the change in the Gibbs free energy ( $\Delta G$ ). Most of these folding studies are typically done using chemical denaturants acting on untethered proteins (i. e. in solution). However, a considerable number of cytoskeletal and extracellular matrix proteins, as well as unfoldase substrates from chaperones, translocases and protein degradation machines, are likely to be subjected to mechanical forces and are tethered (Fig. 1). Thus, for those proteins at least, SMFS experiments (providing that the pulling geometry is the right one) may more closely mimic the physiological conditions in which they function in the cell.

Furthermore, the reaction coordinate in SMFS experiments has a well-defined physical meaning (i. e., protein length) and it is a “natural” one for some mechanical proteins. This is in clear contrast with the less well physically defined kinetic “ $m$ -values” used in chemical folding experiments. The  $m$ -value is defined as the derivative of the natural logarithm of the folding, or unfolding, rate constant with respect to the denaturant concentration and it measures the sensitivity of the rate of the process to denaturant concentration, being generally interpreted as a measure of the change in solvent exposure of the lateral chains of amino acid residues.

Thermodynamic comparisons of the  $\Delta G$  of the process between both methods should in principle give identical results, if the entropic contribution of tethering the ends of the molecule is properly corrected for. At equilibrium, from a single molecule, it is possible to determine  $\Delta G$  (which as a state function, it depends only on the initial and final stages of the process and can be obtained by integrating a reversible force-extension curve), the equilibrium constant, the reaction kinetics and their dependence on force [19]. Most single-molecule unfolding

experiments in proteins are performed under nonequilibrium conditions, where the molecule is pulled or relaxed at a faster rate than its spontaneous rate of equilibration (i. e. its molecular relaxation time). Consequently, a marked hysteresis is apparent (i. e. the extension and relaxation curves do not overlap) and at the same time the unfolding forces become speed dependent (Fig. 3). This indicates that not all the mechanical work carried out during the unfolding reaction is converted into a change in the free energy of the molecule (i. e., the efficiency of the process is less than 100%). It has recently been demonstrated that it is possible to recover the free energy of unfolding RNA molecules not only from near-equilibrium conditions [72] but also in far-from-equilibrium conditions [36]. The latter method is a promising tool to extract the equilibrium free energy of protein unfolding, since most protein unfolding reactions occur far from equilibrium.

In contrast, the kinetics of the reaction depends on the pathway and SMFS unfolding experiments impose a reaction coordinate to the molecule distinct to that of the bulk (chemical) experiments. Force acts along a single dimension in specific regions of the protein, typically the N- and C-termini, while conventional denaturants have a more global effect. As a result, the kinetic parameters of unfolding obtained by both methods may differ. In fact, a poor correlation was found between mechanical and chemical unfolding kinetic rates (a measure of kinetic stability) in a recent survey of 19 proteins [27]. In contrast, there was a remarkable agreement between these unfolding rate constants for modules I27 and I28 from the distal I-band region of titin [25,68]. This is noteworthy and raises interesting questions about the mechanical design of these particular domains. It has been shown that the A'G region is the only region in the I27 fold that is critical in both pathways and it is responsible for kinetic stability in both cases [44,69]. This may explain why the two unfolding rates are so close in this module.

While unfolding rates along the force-unfolding pathway can be easily estimated, refolding rates along the force-folding pathway cannot usually be measured because the high forces applied often prevent the folding of proteins. Using the regular length-clamp mode it has been possible to measure refolding rates during relaxation for several proteins, such as I27, projectin, ankyrin B, myosinII and filamin [18,25,67,104,105]. For example, ankyrin B repeats were shown to refold against force (20–25 pN) with complete reversibility, which makes them true elastic enthalpic springs [67]. The myosin II tail (a dimer that forms a coiled coil supramolecular structure) is also an elastic protein structure able to refold against forces of up to 30 pN, although the transition presented a lit-

tle hysteresis, indicating that the process is not fully reversible [104].

Finally, force-clamp SMFS has been used to directly examine the mechanical folding pathways of I27, ubiquitin and projectin molecules [18,42,47,123] (Fig. 2c). In these experiments, the protein is first unfolded and extended at a high force and then relaxed at lower forces so that refolding can be monitored by measuring changes in the end-to-end length of the protein. Under these conditions, folding is marked by large fluctuations in the end-to-end length of the protein, which have been interpreted as folding of the chain through many continuous steps. By controlling the end-to-end length of a single protein with subnanometer resolution these studies provide us with a new perspective on how to analyze protein folding trajectories (Fig. 8). It should be noted that refolding using the force-clamp mode may proceed along different pathways than those in the absence of such restraints, as shown for ubiquitin using a Go-like model [30]. Finally, optical tweezers (that use lower spring constants and hence, correspondingly lower loading rates than AFM) have been used to study the mechanical unfolding/folding of RNase H and maltose binding protein [8,28]. These studies are providing new molecular insights into folding intermediates, the energy landscape of the process, and how chaperone interactions affect protein folding pathways at the single molecule level.

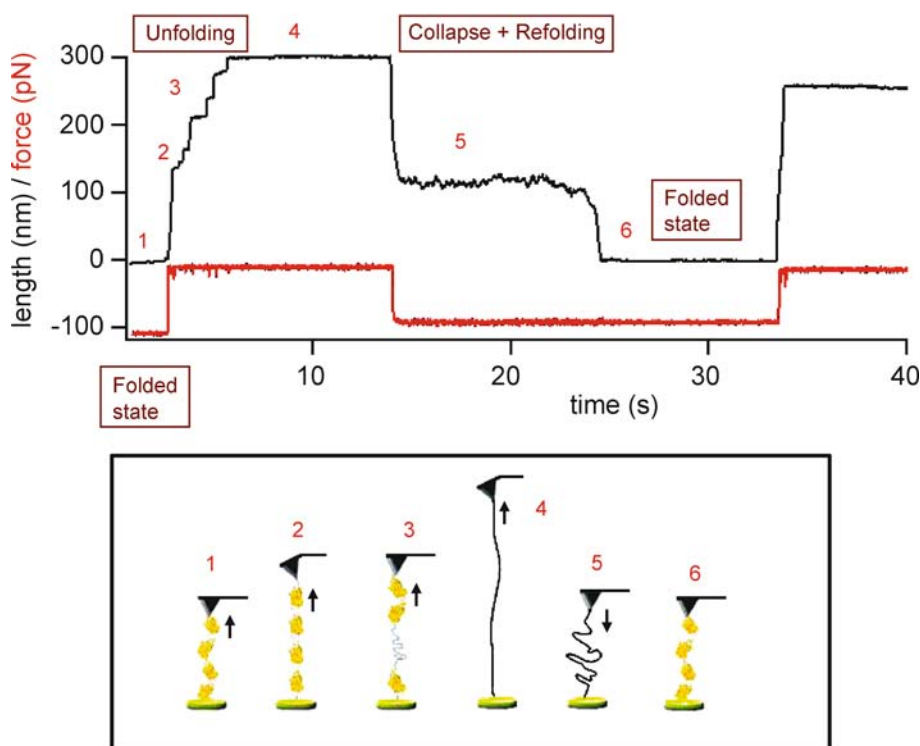
## Future Directions

It has been just a decade since the first individual protein molecule was pulled by SMFS. This impressive feat has since grown into a new discipline providing a wealth of information on the molecular elasticity of proteins, a fundamental property for many biological processes. This new methodology is unveiling the mechanical properties of more and more proteins, and it is providing new insights into the problem of protein folding.

However, only a few protein folds have been analyzed to date and many improvements are still required, such as more specific functionalization methods for immobilization of soluble proteins in order to improve the efficiency and sample control of these experiments. These immobilization methods should ideally be compatible with a *quasi*-simultaneous imaging of proteins in the same sample [119]. We also need single-molecule reporters to conduct more reliable studies on intermolecular interactions between proteins.

It would be also very interesting to use optical tweezers and the biomembrane force probe instrument (which allow even lower loading rates to be applied) on some of the proteins already studied by AFM and to compare the





**Protein Mechanics at the Single-Molecule Level, Figure 8**

Collapse of unfolded titin-like domains under force (mechanical refolding). A titin like protein (projectin) molecule is first unfolded and extended at a high force (97 pN) using force-clamp SMFS. We observed 10 unfolding events. There was an initial large step elongation of  $\sim 100$  nm upon application of force and this initial phase most likely corresponds to the length of the folded polypeptide chain plus a few already unfolded domains. Then after  $\sim 12$  s the protein was relaxed to a force of 15 pN and before the protein reached its fully collapsed state there was a dramatic increase in the noise level with length fluctuations of up to 10 nm peak-to-peak. The source of this noise is not clear, but the phenomenon may reflect the transient formation of secondary structures or intermediate folded conformations. There are three main phases: i) a fast phase ( $< 200$  ms) corresponding to the elastic recoil of the unfolded polypeptide chain and accounting for  $\sim 60\%$  of the unfolded length of the protein; ii) a slow phase ( $\sim 1\text{--}8$  nm/s) characterized by large fluctuations in end-to-end length (up to 10 nm in this example); and iii) again a fast phase ( $> 100$  nm/s) that corresponds to the final collapse of the polypeptide chain to its folded length. In order to test whether the domains are folded, the protein was unfolded by applying a second stretching pulse to 97 pN after 30 s. This experiment demonstrates that titin domains can refold under force; this suggests that titin-like proteins could function according to a folding-based-spring mechanism. (After [18])

results obtained at these low loading rates. For the sake of comparison researchers should also follow a standard set of experimental conditions for SMFS experiments [12,97], as recently proposed in the field of chemical protein folding [78]. Moreover, there is a need for a sensor that could report forces inside the living cell and important advances along these lines were recently reported [57,99,108].

Single-molecule mechanical techniques are still in their infancy, but they are maturing fast. These techniques are providing us more and more fundamental information on the structure and function of proteins. Accordingly, they are becoming an indispensable tool to understand how proteins fold and work in real time. With the new-found capacity to manipulate and look at the “secret life” of single molecules, we should be prepared for many sur-

prises from the mechanochemistry of proteins. Through the information unveiled by these techniques we are entering a new and exciting time in biology which, in combination with the knowledge generated in this proteomic era, is likely to move us closer to understanding the logic behind protein design.

### Acknowledgments

This work was funded by grants from the Spanish Ministry of Science and Education (BIO2007-67116), the Consejería de Educación of the Madrid Community (S-0505/MAT/0283), and the Spanish Research Council (200620F00) to M.C.-V., from the NIH (R01DK073394), the John Sealy Memorial Endowment Fund for Biomedical

Research, and the Polycystic Kidney Foundation (116a2r) to A.F.O., and from the Ministry of Science and Higher Education (N N202 0852 33) to M.C. We apologize to all researchers whose pioneering work was not cited due to limitations of space.

## Bibliography

### Primary Literature

1. Abe H, Go N (1981) Noninteracting local-structure model of folding and unfolding transition in globular proteins. II, Application to two-dimensional lattice proteins. *Biopolymers* 20:1013–1031
2. Abu-Lail NI, Ohashi T, Clark RL, Erickson HP, Zauscher S (2005) Understanding the elasticity of fibronectin fibrils: unfolding strengths of FN-III and GFP domains measured by single-molecule force spectroscopy. *Matrix Biol* 25:175–184
3. Ainavarapu SR, Li L, Badilla CL, Fernandez JM (2005) Ligand binding modulates the mechanical stability of dihydrofolate reductase. *Biophys J* 89:3337–3344
4. Ainavarapu SR, Bruijic J, Huang HH, Wiita AP, Lu H, Li L, Walther KA, Carrion-Vazquez M, Li H, Fernandez JM (2007) Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. *Biophys J* 92:225–233
5. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 92:291–294
6. Bao G, Suresh S (2003) Cell and molecular mechanics of biological materials. *Nat Mater* 2:715–725
7. Batey S, Randles LG, Steward A, Clarke J (2005) Cooperative folding in a multi-domain protein. *J Mol Biol* 349:1045–1059
8. Bechtluft P, van Leeuwen RG, Tyreman M, Tomkiewicz D, Nouwen N, Tepper HL, Driessen AJ, Tans SJ (2007) Direct observation of chaperone-induced changes in a protein folding pathway. *Science* 318:1458–1461
9. Bell GI (1978) Models for the specific adhesion of cells to cells. *Science* 200:618–627
10. Best RB, Li B, Steward A, Daggett V, Clarke J (2001) Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation. *Biophys J* 81:2344–2356
11. Best RB, Fowler SB, Toca-Herrera JL, Clarke J (2002) A simple method for proving the mechanical unfolding pathway of proteins in detail. *Proc Natl Acad Sci USA* 99:12143–12148
12. Best RB, Brockwell DJ, Toca-Herrera JL, Blake AW, Smith DA, Radford SE, Clarke J (2003) Force mode atomic force microscopy as a tool for protein folding studies. *Anal Chim Acta* 479:87–105
13. Best RB, Fowler SB, Toca-Herrera JL, Steward A, Paci E, Clarke J (2003) Mechanical unfolding of a titin Ig domain: Structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *J Mol Biol* 330:867–877
14. Bhasin N, Carl P, Harper S, Feng G, Lu H, Speicher DW, Discher DE (2004) Chemistry on a single protein, vascular cell adhesion molecule-1, during forced unfolding. *J Biol Chem* 279:45865–45874
15. Binnig G, Quate CF, Gerber C (1986) Atomic force microscope. *Phys Rev Lett* 56:930–933
16. Brockwell DJ, Beddard GS, Clarkson J, Zinober RC, Blake AW, Trinick J, Olmsted PD, Smith DA, Radford SE (2002) The effect of core destabilization on the mechanical resistance of I27. *Biophys J* 83:458–472
17. Brockwell DJ, Paci E, Zinober RC, Beddard GS, Olmsted PD, Smith DA, Perham RN, Radford SE (2003) Pulling geometry defines the mechanical resistance of a beta-sheet protein. *Nat Struct Biol* 10:731–737
18. Bullard B, Garcia T, Benes V, Leake MC, Linke WA, Oberhauser AF (2006) The molecular elasticity of the insect flight muscle proteins projectin and kettin. *Proc Natl Acad Sci USA* 103:4451–4456
19. Bustamante C, Chemla YR, Forde NR, Izhaky D (2004) Mechanical processes in biochemistry. *Annu Rev Biochem* 73:705–748
20. Bustamante C, Macosko JC, Wuite GJ (2000) Grabbing the cat by the tail: Manipulating molecules one by one. *Nat Rev Mol Cell Biol* 1:130–136
21. Cao Y, Li H (2007) Polyprotein of GB1 is an ideal artificial elastomeric protein. *Nat Mater* 6:109–114
22. Carl P, Kwok CH, Manderson G, Speicher DW, Discher DE (2001) Forced unfolding modulated by disulfide bonds in the Ig domains of a cell adhesion molecule. *Proc Natl Acad Sci USA* 98:1565–1570
23. Carrion-Vazquez M, Li H, Lu H, Marszalek PE, Oberhauser AF, Fernandez JM (2003) The mechanical stability of ubiquitin is linkage dependent. *Nat Struct Biol* 10:738–743
24. Carrion-Vazquez M, Marszalek PE, Oberhauser AF, Fernandez JM (1999) Atomic force microscopy captures length phenotypes in single proteins. *Proc Nat Acad Sci USA* 96:11288–11292
25. Carrion-Vazquez M, Oberhauser AF, Fowler SB, Marszalek PE, Broedel SE, Clarke J, Fernandez JM (1999) Mechanical and chemical unfolding of a single protein: A comparison. *Proc Nat Acad Sci USA* 96:3694–3699
26. Carrion-Vazquez M, Oberhauser AF, Fisher TE, Marszalek PE, Li H, Fernandez JM (2000) Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. *Prog Biophys Mol Biol* 74:63–91
27. Carrión-Vázquez M, Oberhauser AF, Díez H, Hervás R, Oroz J, Fernández J, Martínez-Martín D (2006) Protein nanomechanics, as studied by AFM single-molecule force spectroscopy. In: Arrondo JLR, Alonso A (eds) *Advanced Techniques in Biophysics*. Springer, Berlin, pp 163–245
28. Cecconi C, Shank EA, Bustamante C, Marqusee S (2005) Direct observation of the three-state folding of a single protein molecule. *Science* 309:2057–2060
29. Cieplak M (2005) Mechanical stretching of proteins: calmodulin and titin. *Physica A* 352:28–42
30. Cieplak M, Szymczak P (2006) Protein folding in a force clamp. *J Chem Phys* 124:194901
31. Cieplak M, Hoang TX, Robbins MO (2002) Folding and stretching in a Go-like model of titin. *Proteins* 49:114–124
32. Cieplak M, Hoang TX, Robbins MO (2004) Thermal effects in stretching of Go-like models of titin and secondary structures. *Proteins: Struct Funct Bio* 56:285–297
33. Cieplak M, Filipek S, Janovjak H, Krzysko KA (2006) Pulling single bacteriorhodopsin out of a membrane: Comparison of simulation and experiment. *Biochim Biophys Acta* 1758:537–544

34. Clausen-Schaumann H, Seitz M, Krautbauer R, Gaub HE (2000) Force spectroscopy with single bio-molecules. *Curr Opin Chem Biol* 4:524–530
35. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* 298:937–953
36. Collin D, Ritort F, Jarzynski C, Smith SB, Tinoco I Jr, Bustamante C (2005) Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature* 437:231–234
37. Dietz H, Rief M (2004) Exploring the energy landscape of GFP by single-molecule mechanical experiments. *Proc Natl Acad Sci USA* 101:16192–16197
38. Dietz H, Rief M (2006) Protein structure by mechanical triangulation. *Proc Natl Acad Sci USA* 103:1244–1247
39. Dietz H, Berkemeier F, Bertz M, Rief M (2006) Anisotropic deformation response of single protein molecules. *Proc Natl Acad Sci USA* 103:12724–12728
40. Evans E, Ritchie K (1997) Dynamic strength of molecular adhesion bonds. *Biophys J* 72:1541–1555
41. Evans E, Ritchie K, Merkel R (1995) Sensitive force technique to probe molecular adhesion and structural linkages at biological interfaces. *Biophys J* 68:2580–2587
42. Fernandez JM, Li H (2004) Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science* 303:1674–1678
43. Fersht AR, Matouschek A, Serrano L (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* 224:771–782
44. Forman JR, Clarke J (2007) Mechanical unfolding of proteins: insights into biology, structure and folding. *Curr Opin Struct Biol* 17:58–66
45. Fowler SB, Best RB, Toca-Herrera JL, Rutherford TJ, Steward A, Paci E, Karplus M, Clarke J (2002) Mechanical unfolding of a titin Ig domain: Structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. *J Mol Biol* 322:841–849
46. Gao M, Sotomayor M, Villa E, Lee EH, Schulten K (2006) Molecular mechanisms of cellular mechanics. *Phys Chem Chem Phys* 8:3692–3706
47. Garcia-Manes S, Brujic J, Badilla CL, Fernandez JM (2007) Force-clamp spectroscopy of single-protein monomers reveals the individual unfolding and folding pathways of I27 and ubiquitin. *Biophys J* 93:2436–2446
48. Geiger B, Bershadsky A (2002) Exploring the neighborhood: adhesion-coupled cell mechanosensors. *Cell* 110:139–142
49. Giannone G, Sheetz MP (2006) Substrate rigidity and force define form through tyrosine phosphatase and kinase pathways. *Trends Cell Biol* 16:213–223
50. Grubmüller H, Heymann B, Tavan P (1996) Ligand binding: Molecular mechanics calculation of the streptavidin biotin rupture force. *Science* 271:997–999
51. Hinterdorfer P (2002) Molecular recognition studies using the atomic force microscope. *Methods Cell Biol* 68:115–139
52. Hoang TX, Cieplak M (2000) Molecular dynamics of folding of secondary structures in Go-like models of proteins. *J Chem Phys* 112:6851–6862
53. Howard J (2001) *Mechanics of Motor Proteins and the Cytoskeleton*. Sinauer Associates, Sunderland
54. Huang L, Kirmizialtin S, Makarov DE (2005) Computer simulations of the translocation and unfolding of a protein pulled mechanically through a pore. *J Chem Phys* 123:124903
55. Huxley H, Hanson J (1954) Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation. *Nature* 173:973–976
56. Hyeon CB, Thirumalai D (2003) Can energy landscape roughness of proteins and RNA be measured by using mechanical unfolding experiments? *Proc Natl Acad Sci USA* 100:10249–10253
57. Johnson CP, Tang HY, Carag C, Speicher DW, Discher DE (2007) Forced unfolding of proteins within cells. *Science* 317:663–666
58. Junker JP, Hell K, Schlierf M, Neupert W, Rief M (2005) Influence of substrate binding on the mechanical stability of mouse dihydrofolate reductase. *Biophys J* 89:L46–48
59. Karanicolas J, Brooks CL III (2002) The origins of asymmetry in the folding transition states of protein L and protein G. *Prot Sci* 11:2351–2361
60. Kedrov A, Janovjak H, Sapra KT, Müller DJ (2007) Deciphering molecular interactions of native membrane proteins by single-molecule force spectroscopy. *Annu Rev Biophys Biomol Struct* 36:233–260
61. Kellermayer MS (2005) Visualizing and manipulating individual protein molecules. *Physiol Meas* 26:R119–R153
62. Kellermayer MS, Smith SB, Granzier HL, Bustamante C (1997) Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science* 276:1112–1116
63. Klimov DK, Thirumalai D (2000) Native topology determines force-induced unfolding pathways in globular proteins. *Proc Natl Acad Sci USA* 97:7254–7259
64. Law R, Carl P, Harper S, Dalhaimer P, Speicher DW, Discher DE (2003) Cooperativity in forced unfolding of tandem spectrin repeats. *Biophys J* 84:533–544
65. Leake MC, Grutzner A, Kruger M, Linke WA (2006) Mechanical properties of cardiac titin’s N2B-region by single-molecule atomic force spectroscopy. *J Struct Biol* 155:263–272
66. Lee C-K, Wang Y-M, Huang L-S, Lin S (2007) Atomic force microscopy: Determination of unbinding force, off rate and energy barrier for protein-ligand interaction. *Micron* 38:446–461
67. Lee G, Abdi K, Jiang Y, Michael P, Bennett V, Marszalek PE (2006) Nanospring behaviour of ankyrin repeats. *Nature* 440:246–249
68. Li H, Oberhauser AF, Fowler SB, Clarke J, Fernandez JM (2000) Atomic force microscopy reveals the mechanical design of a modular protein. *Proc Natl Acad Sci USA* 97:6527–6531
69. Li H, Carrion-Vazquez M, Oberhauser AF, Marszalek PE, Fernandez JM (2000) Point mutations alter the mechanical stability of immunoglobulin modules. *Nat Struct Biol* 7:1117–1120
70. Li H, Linke WA, Oberhauser AF, Carrion-Vazquez M, Kerkvliet JG, Lu H, Marszalek PE, Fernandez JM (2002) Reverse engineering of the giant muscle protein titin. *Nature* 418:998–1002
71. Li H, Oberhauser AF, Redick SD, Carrion-Vazquez M, Erickson HP, Fernandez JM (2001) Multiple conformations of PEVK proteins detected by single-molecule techniques. *Proc Natl Acad Sci USA* 98:10682–10686
72. Liphardt J, Dumont S, Smith SB, Tinoco I Jr, Bustamante C (2002) Equilibrium information from nonequilibrium mea-

- surements in an experimental test of Jarzynski's equality. *Science* 296:1832–1853
73. Lu H, Israilewitz B, Krammer A, Vogel V, Schulten K (1998) Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J* 75:662–671
  74. Makarov DE, Hansma PK, and Metiu H (2001) Kinetic Monte Carlo simulation of titin unfolding. *J Chem Phys* 114:9663–9673
  75. Marszalek PE, Lu H, Li H, Carrion-Vazquez M, Oberhauser AF, Schulten K, Fernandez JM (1999) Mechanical unfolding intermediates in titin modules. *Nature* 402:100–103
  76. Maruyama K, Kimura S, Ohashi K, Kuwano Y (1981) Connectin, an elastic protein of muscle. Identification of "titin" with connectin. *J Biochem (Tokyo)* 89:701–709
  77. Matouschek A (2003) Protein unfolding-an important process *in vivo*? *Curr Opin Struct Biol* 13:98–109
  78. Maxwell KL, Wildes D, Zarrine-Afsar A, de los Rios MA, Brown AG, et al (2005) Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci* 14:602–616
  79. Mehta AD, Rief M, Spudich JA (1999) Biomechanics, one molecule at a time. *J Biol Chem* 274:14517–14520
  80. Merkel R (2001) Force spectroscopy on single passive biomolecules and single biomolecular bonds. *Phys Rep* 346:344–385
  81. Miller E, Garcia T, Hultgren S, Oberhauser AF (2006) The mechanical properties of *E. coli* type 1 pili measured by atomic force microscopy techniques. *Biophys J* 91:3848–3856
  82. Müller DJ, Engel A (2007) Atomic force microscopy and spectroscopy of native membrane proteins. *Nat Protoc* 2:2191–2197
  83. Neher E, Sakmann B (1976) Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature* 260:799–802
  84. Neuman KC, Lionnet T, Allemand J-F (2007) Single-molecule micromanipulation techniques. *Annu Rev Mater Res* 37:33–67
  85. Ng SP, Rounsevell RWS, Steward A, Geierhaas CD, Williams PM, Paci E, Clarke J (2005) Mechanical unfolding of TNfn3: The unfolding pathway of a fnIII domain probed by protein engineering, AFM and MD simulation. *J Mol Biol* 350:776–789
  86. Oberhauser AF, Carrión-Vázquez M (2008) Mechanical biochemistry of proteins one molecule at a time. *J Biol Chem* 283:6617–6621
  87. Oberhauser AF, Marszalek PE, Erickson HP, Fernandez JM (1998) The molecular elasticity of tenascin, an extracellular matrix protein. *Nature* 393:181–185
  88. Oberhauser AF, Marszalek PE, Carrion-Vazquez M, Fernandez JM (1999) Single protein misfolding events captured by atomic force microscopy. *Nat Struct Biol* 6:1025–1028
  89. Oberhauser AF, Badilla-Fernandez C, Carrion-Vazquez M, Fernandez JM (2002) The mechanical hierarchies of fibronectin observed with single-molecule AFM. *J Mol Biol* 319:433–447
  90. Perez-Jimenez R, Garcia-Manyes S, Ainaravaru SR, Fernandez JM (2006) Mechanical unfolding pathways of the enhanced yellow fluorescent protein revealed by single molecule force spectroscopy. *J Biol Chem* 281:40010–40014
  91. Prakash S, Matouschek A (2004) Protein unfolding in the cell. *Trends Biochem Sci* 29:593–600
  92. Qian F, Wei W, Germino G, Oberhauser A (2005) The nanomechanics of polycystin-1 extracellular region. *J Biol Chem* 280:40723–40730
  93. Rief M, Grubmüller H (2002) Force spectroscopy of single biomolecules. *Chemphyschem* 3:255–261
  94. Rief M, Gautel M, Oesterhelt F, Fernandez JM, Gaub HE (1997) Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* 276:1109–1112
  95. Rief M, Pascual J, Saraste M, Gaub HE (1999) Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *J Mol Biol* 286:553–561
  96. Ritort F (2006) Single-molecule experiments in biological physics: Methods and applications. *J Phys: Condens Matter* 18:R531–R583
  97. Rounsevell R, Forman JR, Clarke J (2004) Atomic force microscopy: mechanical unfolding of proteins. *Methods* 34:100–111
  98. Samorì B, Zuccheri G, Baschieri R (2005) Protein unfolding and refolding under force: Methodologies for nanomechanics. *Chemphyschem* 6:29–34
  99. Sarkar A, Robertson RB, Fernandez JM (2004) Simultaneous atomic force microscope and fluorescence measurements of protein unfolding using a calibrated evanescent wave. *Proc Natl Acad Sci USA* 101:12882–12886
  100. Sarkar A, Caamano S, Fernandez JM (2005) The elasticity of individual titin PEVK exons measured by single molecule atomic force microscopy. *J Biol Chem* 280:6261–6264
  101. Sato T, Esaki M, Fernandez JM, Endo T (2005) Comparison of the protein-unfolding pathways between mitochondrial protein import and atomic-force microscopy measurements. *Proc Natl Acad Sci USA* 102:17999–18004
  102. Sauer RT, Bolon DN, Burton BM, Burton RE, Flynn JM, Grant RA, Hersch GL, Joshi SA, Kenniston JA, Levchenko I, Neher SB, Oakes E, Siddiqui SM, Wah DA, Baker TA (2004) Sculpting the proteome with AAA(+) proteases and disassembly machines. *Cell* 119:9–18
  103. Schlierf M, Li H, Fernandez JM (2004) The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. *Proc Natl Acad Sci USA* 101:7299–7304
  104. Schwaiger I, Sattler C, Hostetter DR, Rief M (2002) The myosin coiled-coil is a truly elastic protein structure. *Nat Mater* 1:232–235
  105. Schwaiger I, Schleicher M, Noegel AA, Rief M (2005) The folding pathway of a fast-folding immunoglobulin domain revealed by single-molecule mechanical experiments. *EMBO Rep* 6:1–6
  106. Sharma D, Perisic O, Peng Q, Cao Y, Lam C, Lu H, Li H (2007) Single-molecule force spectroscopy reveals a mechanically stable protein fold and the rational tuning of its mechanical stability. *Proc Natl Acad Sci USA* 104:9278–9283
  107. Shtilerman M, Lorimer GH, Englander SW (1999) Chaperonin function: folding by forced unfolding. *Science* 284:822–825
  108. Smith ML, Gourdon D, Little WC, Kubow KE, Eguluz RA, Luna-Morris S, Vogel V (2007) Force-induced unfolding of fibronectin in the extracellular matrix of living cells. *PLoS Biol* 5:e268
  109. Sotomayor M, Schulten K (2007) Single-molecule experiments *in vitro* and *in silico*. *Science* 316:1144–1148
  110. Sulkowska JI, Cieplak M (2007) Mechanical stretching of proteins- A theoretical survey of the Protein Data Bank – a topical review. *J Phys Cond Mat* 19:283201



111. Sulkowska JI, Cieplak M (2008) Stretching to understand proteins-A survey of the Protein Data Bank. *Biophys J* 94:6–13
112. Sulkowska JI, Sulkowski P, Szymczak P, Cieplak M (2008) Tightening of knots in proteins. *Phys Rev Lett* 100:058106
113. Sulkowska JI, Cieplak M (2008) Selection of optimal variants of Go-like models of protein through studies of stretching. *Biophys J* 95:3174–3191
114. Takada S (1999) Go-ing for the prediction of protein folding mechanisms. *Proc Natl Acad Sci USA* 96:11609–11700
115. Tian P, Andricioaei I (2005) Repetitive pulling catalyzes co-translocational unfolding of barnase during import through a mitochondrial pore. *J Mol Biol* 350:1017–1034
116. Trylska J, Tozzini V, McCammon JA (2005) Exploring global motions and correlations in the ribosome. *Biophys J* 89:1455–1463
117. Tskhovrebova L, Trinick J (2003) Titin: properties and family relationships. *Nat Rev Mol Cell Biol* 4:679–789
118. Urry DW, Hugel T, Seitz M, Gaub HE, Sheiba L, Dea J, Xu J, Parker T (2002) Elastin: a representative ideal protein elastomer. *Philos Trans R Soc Lond B Biol Sci* 357:169–184
119. Valbuena A, Oroz J, Vera AM, Gimeno A, Gómez-Herrero J, Carrión-Vázquez M (2007) *Quasi*-simultaneous imaging/pulling analysis of polyprotein molecules by AFM. *Rev Sci Instrum* 78:113707
120. Veitschan T, Klimov D, Thirumalai D (1997) Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding Des* 2:1–22
121. Vogel V (2006) Mechanotransduction involving multimodular proteins: converting force into biochemical signals. *Annu Rev Biophys Biomol Struct* 35:459–488
122. Vogel V, Sheetz M (2006) Local force and geometry sensing regulate cell functions. *Nat Rev Mol Cell Biol* 7:265–275
123. Walther KA, Grater F, Dougan L, Badilla CL, Berne BJ, Fernandez JM (2007) Signatures of hydrophobic collapse in extended proteins captured with force spectroscopy. *Proc Natl Acad Sci USA* 104:7916–7921
124. Wang K, McClure J, Tu A (1979) Titin: major myofibrillar components of striated muscle. *Proc Natl Acad Sci USA* 76:3698–3702
125. Weisel JW, Shuman H, Litvinov RI (2003) Protein-protein unbinding induced by force: Single-molecule studies. *Curr Opin Struct Biol* 13:227–235
126. West DK, Brockwell DJ, Olmsted PD, Radford SE, Paci E (2006) Mechanical resistance of proteins explained using simple molecular models. *Biophys J* 90:287–297
127. West DK, Brockwell DJ, Paci E (2006) Prediction of the translocation kinetics of a protein from its mechanical properties. *Biophys J* 91:L51–53
128. Wiita AP, Ainaravapu SR, Huang HH, Fernandez JM (2006) Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. *Proc Natl Acad Sci USA* 103:7222–7227
129. Wiita AP, Perez-Jimenez R, Walther KA, Grater F, Berne BJ, Holmgren A, Sanchez-Ruiz JM, Fernandez JM (2007) Probing the chemistry of thioredoxin catalysis with force. *Nature* 450:124–127
130. Wilcox AJ, Choy J, Bustamante C, Matouschek A (2005) Effect of protein structure on mitochondrial import. *Proc Natl Acad Sci USA* 102:15435–15440
131. Yang G, Cecconi C, Baase WA, Vetter IR, Breyer WA, Haack JA, Matthews BW, Dahlquist FW, Bustamante C (2000) Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme. *Proc Natl Acad Sci USA* 97:139–144
132. Zhuang X, Rief M (2003) Single-molecule folding. *Curr Opin Struct Biol* 13:88–97
133. Zlatanova J, Lindsay SM, Leuba SH (2000) Single molecule force spectroscopy in biology using the atomic force microscope. *Prog Biophys Mol Biol* 74:37–61
134. Zlatanova J, van Holde K (2006) Single-molecule biology: What is it and how does it work? *Mol Cell* 24:317–329

## Books and Reviews

- Boal D (2002) *Mechanics of the Cell*. Cambridge University Press, Cambridge
- Brande C, Tooze J (1999) *Introduction to Protein Structure*, 2nd edn. Garland Publishing, New York
- Goodsell DS (2004) *Bionanotechnology: Lessons from Nature*. Wiley, Hoboken
- Grosberg AY, Khokhlov AR (1997) *Giant Molecules: Here, There and Everywhere*. Academic Press, San Diego
- Grubmüller H, Schulten K (eds) (2007) *Advances in molecular dynamics simulations*. *J Struct Biol* 157(special issue):443–615
- Leuba SH, Zlatanova J (2000) *Biology at the Single-Molecule Level*. Pergamon Press, Oxford
- Wainwright SA, Biggs WD, Currey JD, Gosline JM (1976) *Mechanical Design in Organisms*. Princeton University Press, Princeton

## Public Policy, System Dynamics Applications to

DAVID F. ANDERSEN<sup>1</sup>, ELIOT RICH<sup>2</sup>,

RODERICK MACDONALD<sup>3</sup>

<sup>1</sup> Rockefeller College of Public Affairs and Policy,  
University at Albany, Albany, USA

<sup>2</sup> School of Business, University at Albany, Albany, USA

<sup>3</sup> Initiative for System Dynamics in the Public Sector,  
University at Albany, Albany, USA

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Medical Malpractice: A System Dynamics  
and Public Policy Vignette](#)

[What Is System Dynamics Modeling?](#)

[How Is System Dynamics Used to Support Public Policy  
and Management?](#)

[System Dynamics and Models:](#)

[A Range of Analytic Scope and Products](#)

[What Are the Arenas in Which System  
Dynamics Models Are Used?](#)



What Are some of the Substantive Areas  
Where System Dynamics Has Been Applied?  
Evaluating the Effectiveness of System Dynamics Models  
in Supporting the Public Policy Process  
Summary: System Dynamics – A Powerful Tool  
to Support Public Policy  
Future Directions  
Bibliography

## Glossary

**Causal loop diagram** A diagrammatic artifact that captures the causal model and feedback structure underlying a problem situation. Commonly used as a first-cut tool to identify major stakeholder concerns and interactions. These diagrams are often precursors to formal models.

**Dynamic modeling** Formal examination of the behavior of a system over time. Contrast with point-estimation, which attempts to predict an average outcome.

**Feedback** A relationship where two or more variables are linked over time so that the influence of one variable on a second will later affect the state of the first. If the influence is such as to increase the state of the first over time, the feedback is termed reinforcing. If the influence is such as to decrease the state of the first, it is termed balancing.

**Formal model** The representation of a system structure in mathematical form. Contrast with causal model, which represents structure without the underlying mathematics.

**Mental model** The representation of a problem's structure as possessed by an expert in a particular domain. Mental models are often intangible until explicated by the expert.

**Public policy** Any and all actions or non-actions, decisions or non-decisions taken by government, at all levels, to address problems. These actions, non-actions, decisions or non-decisions are implemented through laws, regulations and the allocation of resources.

**Group model building (GMB)** An approach to problem definition that asks multiple experts and major stakeholders to provide collective insights into the structure and behavior of a system through facilitated exercises and artifacts. GMB is often used to explicate the contrasting mental models of stakeholders.

**Stakeholder** An individual or group that has significant interest or influence over a policy problem.

**System dynamics** An analytic approach to problem definition and solution that focuses on endogenous variables linked through feedback, information and mate-

rial delays, and non-linear relationships. The structure of these linkages determines the behavior of the modeled system.

## Definition of the Subject

System dynamics is an approach to problem understanding and solution. It captures the complexity of real-world problems through the explication of feedback among endogenous variables. This feedback, and the delays that accompany it, often drive public sector programs towards unanticipated or unsatisfactory results. Through formal and informal modeling, System Dynamics-based analysis explicates and opens these feedback structures to discussion, debate and consensus building necessary for successful public sector policymaking.

## Introduction

In the 50 years since its founding, System Dynamics has contributed to public policy thought in a number of areas. Major works, such as *Urban Dynamics* [35] and *Limits to Growth* [61] have sparked controversy and debate. Other works in the domains of military policy, illegal drugs, welfare reform, health care, international development, and education have provided deep insight into complex social problems. The perspective of System Dynamics, with its emphasis on feedback, changes over time, and the role of information delays, helps inform policy makers about the intended and unintended consequences of their choices. The System Dynamics method includes a problem-oriented focus and the accommodation of multiple stakeholders, both crucial to the development of sound policy. Through the use of formal simulation, decision makers may also use System Dynamics models to consider the effects of their choices on short- and long-term outcomes. We illustrate this process with real life examples, followed by a review of the features of System Dynamics as they relate to public policy issues. We then describe the conjunction of System Dynamics and Group Model Building as a mechanism for policy ideation and review. We identify some of the historical and current uses of System Dynamics in the public sector, and discuss techniques for evaluating its effects on policy and organizations.

## Medical Malpractice: A System Dynamics and Public Policy Vignette

*The year was 1987 and New York's medical malpractice insurance system was in a state of crisis. Fueled by unprecedented levels of litigation, total settlements were soaring as were the malpractice insurance rates charged to hospitals*

and physicians. Obstetricians stopped taking on new patients. Doctors threatened to or actually did leave the state. Commercial insurance carriers had stopped underwriting malpractice insurance policies, leaving state-sponsored risk pools as the only option. The Governor and the Legislature were under pressure to find a solution and to find it soon. At the center of this quandary was the state's Insurance Department, the agency responsible for regulating and setting rates for the state's insurance pools. The agency's head found himself in just the kind of media hot seat one seeks to avoid in the public service.

An in-house SWAT team of actuaries, lawyers, and analysts had been working to present a fiscally sound and politically viable set of options for the Agency to consider and recommend to the legislature. They had been working with a team of System Dynamics modelers to gain better understanding of the root causes of the crisis. Working as a group, they had laid out a whole-system view of the key forces driving malpractice premiums in New York State. Their simulation model, forged in the crucible of group consensus, portrayed the various options on a "level playing field," each option being analyzed using a consistent set of operating assumptions. One option stood out for its ability to offer immediate malpractice insurance premium relief, virtually insuring a rapid resolution to the current crisis. An actuarial restructuring of future liabilities arising from future possible lawsuits relieved immediate pressure on available reserve funds. Upward pressure on premiums would vanish; a showdown in the legislature would be averted. Obviously, the Commissioner was interested in this option—who would not be?

"But what happens in the later years, after our crisis is solved?" he asked. As the team pored over the simulation model, they found that today's solution sowed the seeds for tomorrow's problems. Ten, fifteen, or maybe more years into the future, the deferred liabilities piled up in the system creating a secondary crisis, quite literally a second crisis caused by the resolution of the first crisis.

"Take that option off the table – it creates an unacceptable future," was the Commissioner's snap judgment. At that moment a politically appointed official had summarily dismissed a viable and politically astute "silver bullet" cure to a current quandary because he was thinking dynamically, considering both short-term and long-term effects of policy.

The fascinating point of the medical malpractice vignette is that the option taken off the table was indeed, in the short run, a "silver bullet" to the immediate crisis. The System Dynamics model projected that the solution's unraveling would occur long after the present Commissioner's career was over, as well as after the elected life span

of the Governor who had appointed him and the legislators whose votes would be needed to implement the solution. His decision did not define the current problem solely in terms of the current constellation of stakeholders at the negotiations, each with their particular interests and points of view. His dynamic thinking posed the current problem as the result of a system of forces that had accumulated in the past. Symmetrically, his dynamic thinking looked ahead in an attempt to forecast what would be the future dynamic consequences of each option. Might today's solution become tomorrow's problem?

This way of thinking supported by System Dynamics modeling invites speculation about long-run versus short-run effects. It sensitizes policy makers to the pressure of future possible stakeholders, especially future generations who may come to bear the burden of our current decisions. It draws attention into the past seeking causes that may be buried at far spatial and temporal distances from current symptoms within the system. It seeks to understand the natural reaction time of the system, the period during which problems emerge and hence over which they need to be solved. System Dynamics-based analysis in the public sector draws analytic attention away from the riveting logic of the annual or biannual budget cycle, often focusing on options that will play themselves out years after current elected officials have left office. Such work is hard to do, but critical if one wants to think in systems terms.

### What Is System Dynamics Modeling?

While other papers in this series may provide a more expanded answer to this basic question, it may be useful to begin this discussion of System Dynamics and public policy with a brief description of what System Dynamics is.

System Dynamics is an approach to policy analysis and design that applies to problems arising in complex social, managerial, economic, or ecological systems [31,33,74,95]. System Dynamics models are built around a particular problem. The problem defines the relevant factors and key variables to be included in the analysis. This represents the model's boundary, which may cross departmental or organizational boundaries. One of the unique advantages of using System Dynamics models to study public policy problems is that assumptions from a variety of stakeholders are explicitly stated, can be tested through simulation, and can be examined in context.

System Dynamics models rely on three sources of information: numerical data, the written database (reports, operations manuals, published works, etc.), and the expert knowledge of key participants in the system [36]. The numerical database of most organizations is very small,

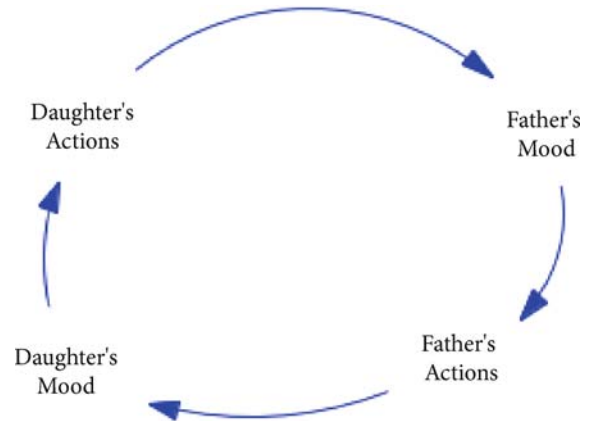
the written database is larger, and the expert knowledge of key participants is vast. System Dynamicists rely on all three sources, with particular attention paid to the expert knowledge of key participants because it is only through such expert knowledge that we have any knowledge of the structure of the system. The explicit capturing of accumulated experience from multiple stakeholders in the model is one of the major differences between System Dynamics models and other simulation paradigms. An understanding of the long term effects of increased vigilance on the crime rate in a community needs to account for the reaction of courts, prisons, and rehabilitation agencies pressed to manage a larger population. This knowledge is spread across experts in several fields, and is not likely to be found in any single computer database. Rather, insight requires a process that makes these factors visible and explicit. For public sector problems, in particular, this approach helps move conflict out of the realm of inter-organizational conflict and towards a problem-solving focus.

Through the use of available data and by using the verbal descriptions of experts to develop mathematical relationships between variables, we expose new concepts and/or previously unknown but significant variables. System Dynamics models are appropriate to problems that arise in closed-loop systems, in which conditions are converted into information that is observed and acted upon, changing conditions that influence future decisions [69].

This idea of a “closed loop” or “endogenous” point of view on a system is really important to all good System Dynamics models. A simple example drawn from everyday life may help better to understand what an endogenous (versus exogenous) point of view means. If a father believes that his teenage daughter is always doing things to annoy him and put him in a bad mood, then he has an exogenous or “open loop” view of his own mood because he is seeing his mood as being controlled by forces outside of or exogenous to his own actions. However, if the father sees that his daughter and her moods are reacting to his own actions and moods while in turn his daughter’s actions shape and define his moods, then this father has an endogenous point of view on his own mood. He understands how his mood is linked in a closed loop with another member of his family. Of course, the father with an endogenous view will be in a better position to more fully understand family dynamics and take actions that can prevent bad moods from spreading within the family.

### Using an SD Model to Develop a Theory

A System Dynamics model represents a theory about a particular problem. Since models in the social sciences



**Public Policy, System Dynamics Applications to, Figure 1**  
Closed loop diagram of fathers and daughters

represent a theory, the most we can hope for from all these models, mental or formal, is that they be useful [94]. System Dynamics models are useful because the mathematical underpinning needed for computer simulation requires that the theory be precise. The process of combining numerical data, written data, and the knowledge of experts in mathematical form can identify inconsistencies about how we think the system is structured and how it behaves over time [38].

In policymaking it is often easy and convenient to blame other stakeholders for the problem state. Often, though, the structure of the system creates the problem by, for example, shifting resources to the wrong recipient or by inclusion of policies that intervene in politically visible but ineffective ways. The use of inclusive SD models educates us by identifying these inconsistencies through an iterative process involving hypotheses about system structure and tests of system behavior. Simulation allows us to see how the complex interactions we have identified work when they are all active at the same time. Furthermore, we can test a variety of policies quickly to see how they play out in the long run. The final result is a model that represents our most insightful and tested theory about the endogenous sources of problem behavior.

### Behavior over Time Versus Forecasts

People who take a systems view of policy problems know that behavior generated by complex organizations cannot be well understood by examining the parts. By taking this holistic view, System Dynamicists capture time delays, amplification, and information distortion as they exist in organizations. By developing computer simulation models that incorporate information feedback, systems modelers seek to understand the internal policies and decisions, and

the external dynamic phenomena that combine to generate the problems observed. They seek to predict dynamic implications of policy, not forecast the values of quantities at a given time in the future.

System Dynamics models are tools that examine the behavior of key variables over time. Historical data and performance goals provide baselines for determining whether a particular policy generates behavior of key variables that is better or worse, when compared to the baseline or other policies. Furthermore, models incorporating rich feedback structure often highlight circumstances where the forces governing a system may change in a radical fashion. For example, in early phases of its growth a town in an arid region may be driven by a need to attract new jobs to support its population. At some future point in time, the very fact of successful growth may lead to a water shortage. Now the search for more water, not more jobs, may be what controls growth in the system. Richardson [69] has identified such phenomena as shifts in loop dominance that provide endogenous explanations for specific outcomes. Simulation allows us to compress time [95] so that many different policies can be tested, the outcomes explained, and the causes that generate a specific outcome can be examined by knowledgeable people working in the system, before policies are actually implemented.

Excellent short descriptions of System Dynamics methodology are found in Richardson [69,70] and Barlas [9]. Furthermore, Forrester's [33] detailed explanation of the field in *Industrial Dynamics* is still relevant, and Richardson and Pugh [74], Roberts et al. [78], Coyle [22], Ford [31], Maani and Cavana [53], Morecroft [65] and Sterman [95] are books that describe the field and provide tools, techniques and modeling examples suitable for the novice as well as for experienced System Dynamics modelers.

### An Application of System Dynamics – The Governor's Office of Regulatory Assistance (GORA) Example

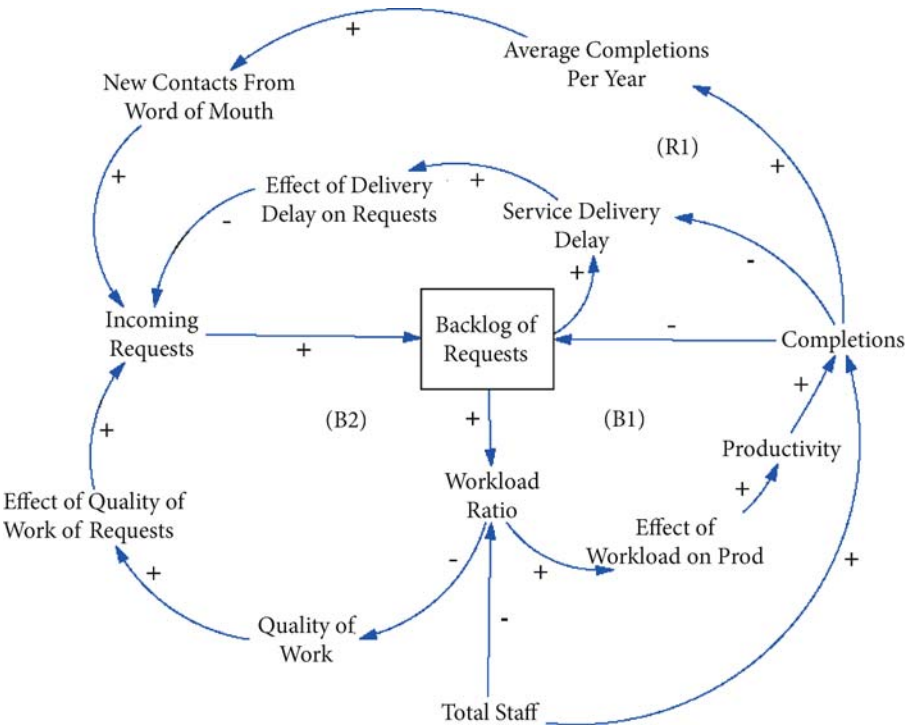
When applied to public policy problems, the “nuts and bolts” of this System Dynamics process consist of identifying the problem, examining the behavior of key variables over time, creating a visualization of the feedback structure of the causes of the problem, and developing a formal simulation model. A second case illustration may assist in understanding the process. The New York State Governor's Office of Regulatory Assistance (GORA) is a governmental agency whose mission it is to provide information about government rules and regulations to entrepreneurs who seek to start up new businesses in the state. The case was described by Andersen et al. [7] and is often used as

a teaching case introducing System Dynamics to public managers.

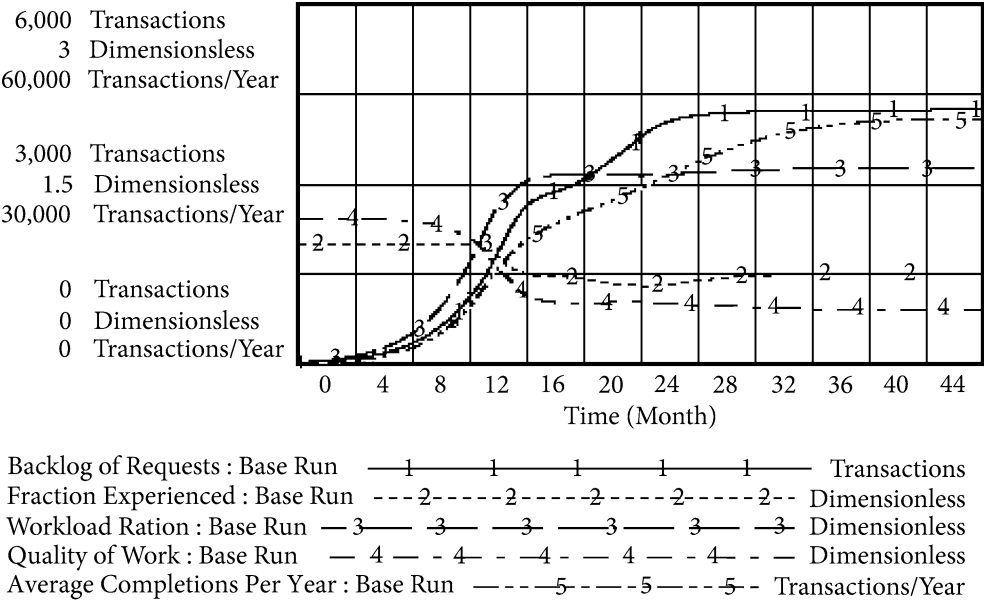
Figure 2 below illustrates three key feedback loops that contribute both to the growth and eventual collapse of citizen service requests at GORA. The reinforcing feedback loop labeled “R1” illustrates how successful completion of citizen orders creates new contacts from word-of-mouth by satisfied citizens which in turn leads to more requests for service coming into the agency. If only this loop were working, a self-reinforcing process would lead to continuing expansion of citizen requests for services at GORA. The balancing loop labeled “B2” provides a balancing effect. As workers within the agency get more and more work to complete, the workload within the agency goes up with one effect being a possible drop in the quality in the work completed. Over time, loop B2 tells a story of how an increased workload can lead to a lower quality of work, with the effect of that lower quality being fewer incoming requests in the future. So over time, too many incoming requests set off a process that limits future requests by driving down quality. Many public managers who have worked with the GORA model find these two simple feedback loops to be realistic and powerful explanations of many of the problems that their agencies face on a day-to-day basis. The full GORA model has many other feedback loops and active variables not shown in the aggregated Fig. 2.

Once all of the variables have been represented by mathematical equations, a computer simulation is able to recreate an over time trajectory possible future values for all of the variables in the model. Figure 3 shows a graph over time of simulated data for key indicators in the GORA case study. The simulation begins when GORA comes into existence to provide services to the public and runs for 48 months. Initially, there is adequate staff and the amount of work to do is low, so the Workload Ratio, shown as part of loops B1 and B2 in the previous figure, is very low. With a low Workload Ratio GORA employees are able to devote additional time to each task they perform and the Quality of Work<sup>1</sup> is thus relatively high. The Backlog of Requests and the Average Completions Per Year begin at 0 and then increase and level off over time to approximately 4,500 and 41,000 respectively. The Fraction Experienced Staff mea-

<sup>1</sup>The Workload Ratio and Quality of Work are normalized variables. This means that they are measured against some predetermined standard. Therefore, when these two variables are equal to 1 they are operating in the desired state. Depending on the definition of the variable, values below or above 1 indicate when they are operating in a desired or undesired state. For example, Quality of Work above 1 indicates that quality is high, relative to the predetermined normal. However, Quality of Work below 1 indicates an undesirable state.



Public Policy, System Dynamics Applications to, Figure 2  
Key feedback loops in a simulation of workflow in the Governor's Office of Regulatory Administration (GORA)



Public Policy, System Dynamics Applications to, Figure 3  
Simulated performance of key variables in the GORA case study



sures what portion of the overall workers are experienced and hence more efficient at doing their jobs. As shown in Fig. 3, the Fraction Experienced begins at 1 and then falls and increases slightly to .75 indicating that GORA is having a harder time retaining experienced staff and is experiencing higher employee turnover. (The full GORA model has a theory of employee burnout and turnover not shown in Fig. 2.)

The combination of the visualization in Fig. 2 with a formal model capable of generating the dynamic output shown in Fig. 3 illustrates the power of System Dynamics modeling for public policy issues. Linking behavior and structure helps stakeholders understand why the behavior of key variables unfolds over time as it does. In the GORA case, the program is initially successful as staff are experienced, are not overworked, and the quality of the services they provide is high. As clients receive services the R1 feedback loop is dominant and this attracts new clients to GORA. However, at the end of the first year the number of clients requesting services begins to exceed the ability of GORA staff to provide the requested services in a timely manner. The Workload Ratio increases, employees are very busy, the Quality of Work falls, and the B2 feedback loop works to limit the number of people seeking services. Furthermore, people are waiting longer to receive services and some are discouraged from seeking services due to the delay. The initial success of the program cannot be sustained and the program settles down into an unsatisfactory situation where the Workload Ratio is high, Quality of Work is low, clients are waiting longer for services and staff turnover is high as indicated by the Fraction Experienced.

The model tells a story of high performance expectations, initial success and later reversal, all explained endogenously. Creating and examining the simulation helps managers consider possible problems before they occur – before staff are overtaxed, before turnover climbs, and before the agency has fallen behind. Having a model to consider compresses time and provides the opportunity for a priori analysis. Finally, having a good model can provide managers with a test bed for asking “what if” questions, allowing public managers to spend simulated dollars and make simulated errors all the while learning how to design better public policies at relatively low cost and without real (only simulated) risk.

### How Is System Dynamics Used to Support Public Policy and Management?

The Medical Malpractice vignette that opened this chapter involving the New York State Commissioner of Insurance

is more fully documented by Reagan-Cirincione et al. [68] and is one of the first published examples of the results of a team of government executives working in a face-to-face group model building session to create a System Dynamics model to support critical policy decisions facing the group. The combined group modeling and simulation approach had a number of positive effects on the policy process. Those positive effects are:

#### Make Mental Models of Key Players Explicit

When the Commissioner drew together his team, the members of this group held different pieces of information and expertise. Much of the most important information was held in the minds, in the mental models, of the Commissioner’s staff, and not in data tabulations. The System Dynamics modeling process made it possible for managers to explicitly represent and manipulate their shared mental models in the form of a System Dynamics simulation model. This process of sharing and aligning mental models, as done during a System Dynamics modeling intervention, is an important aspect of a “learning organization” as emphasized by Senge [87].

#### Create a Formal and Explicit Theory of the Public Policy Situation Under Discussion

The formal model of malpractice insurance contained an explicit and unambiguous theory of how the medical malpractice system in New York State functions. The shared mental models of the client team implied such a formal and model-based theory, but the requirements of creating a running simulation forced the group to be much more explicit and clear about their joint thinking. As the modeling team worked with the group, a shared consensus about how the whole medical malpractice system worked was cast, first into a causal-loop diagram, and later into the equations of the formal simulation model [74,95].

#### Document all Key Parameters and Numbers Supporting the Policy Debate

In addition to creating a formal and explicit theory, the System Dynamics model was able to integrate explicit data and professional experience available to the Department of Insurance. Recording the assumptions of the model in a clear and concise way makes possible review and examination by those not part of the model’s development. Capturing these insights and their derivation provides face validity to the model’s constructs.

Building confidence in the utility of a System Dynamics model for use in solving a public policy problem in-

volves a series of rigorous tests that probe how the model behaves over time as well as how available data, both numerical and tacit structural knowledge, have been integrated and used in the model. Forrester and Senge [39] detail 17 tests for building confidence in a System Dynamics model. Sterman [95] identifies 12 model tests, the purpose or goal of each test, and the steps that modelers should follow in undertaking those tests. Furthermore, Sterman [95] also lists questions that model consumers should ask in order to generate confidence in a model. This is particularly important for public policy issues where the ultimate goal or outcome for different stakeholders may be shared, but underlying assumptions of the stakeholders may be different.

### Create a Formal Model that Stimulates and Answers Key “what if” Questions

Once the formal model was constructed, the Commissioner and his policy team were able to explore “what if” scenarios in a cost-free and risk-free manner. Significant cost overruns in a simulated environment do not drive up real tax rates, nor do they lead to an elected official being voted out of office, nor to an appointed official losing her job. Quite the contrary, a simulated cost overrun or a simulated failed program provides an opportunity to learn how better to implement or manage the program or policy (or to avoid trying to implement the policy). Public managers get to experiment quickly with new policies or programs in a risk-free simulated environment until they “get it right” in the simulated world. Only then should they take the risk of implementation in a high stakes policy environment.

Bringing a complex model to large groups sometimes requires the development of a more elaborate simulation, so that those who were not part of the initial analysis can also derive insight from its results. Iterative development and discussion provides an additional validation of the constructs and conclusions of the model, Zagonel et al. [108] have described a case where local managers responsible for implementing the 1996 federal welfare reform legislation used a simulation model to explore such “what if” futures before taking risks of actual implementation.

Public policy problems are complex, cross organizational boundaries, involve stakeholders with widely different perspectives, and evolve over time. Changes in police procedures and/or resources may have an effect on prison and parole populations many years into the future. Health care policies will determine how resources are allocated at local hospitals and the types of treatments that can be

obtained. Immigration policies in one country may influence the incomes and jobs of people in a second country. Miyakawa [64] has pointed out that public policies are systematically interdependent. Solutions to one problem often create other problems. Increased enforcement of immigration along the U.S. borders has increased the workload of courts [26]. Besides being complex these examples also contain stakeholders with different sets of goals. In solving public policy problems, how diverse stakeholders work out their differences is a key component of successful policy solutions. System Dynamics modeling interventions, and in particular the techniques of group model building [2,6,72,98], provide a unique combination of tools and methods to promote shared understanding by key stakeholders within the system.

### System Dynamics and Models: A Range of Analytic Scope and Products

In the malpractice insurance example, the Commissioner called his advisors into a room to explicitly engage in a group model building session. These formal group model-building sessions involve a specialized blend of projected computer support plus professional facilitation in a face-to-face meeting of public managers and policy analysts. Figure 4 is an illustration of a team of public managers working together in a group model building project. In this photograph, a facilitator is working on a hand drawn view of a simulation model’s structure while projected views of computer output can be used to look at first cut simulation runs or refined images of the model being built by the group. Of course, the key feature of this whole process is facilitated face-to-face conversations between the key stakeholders responsible for the policy decisions being made.

Richardson and Andersen [72], Andersen and Richardson [6], Vennix [98], and Luna-Reyes et al. [52] have provided detailed descriptions of how this kind of group model building process actually takes place. In addition to these group model building approaches, the System Dynamics literature describes five other ways that teams of modelers work with client groups. They are (1) the Reference Group approach [91], (2) the Strategic Forum [75], (3) The stepwise approach [104], (4) strategy dynamics [100,101,102], and (5) the “standard method” of Hines [67].

Some System Dynamics-oriented analyses of public policies completed by groups of public managers and policy analysts stop short of building a formal simulation model. The models produced by Wolstenholme and Coyle [107], Cavana, Boyd and Taylor [14] and the system



**Public Policy, System Dynamics Applications to, Figure 4**

A team of public managers working together to build a System Dynamics model of welfare reform policies

archetypes promoted by Senge [87] have described how these qualitative system mapping exercises, absent a formal running simulation model, can add significant value to a client group struggling with an important public policy problem. The absence of a formal simulation limits the results to a conceptual model, rather than a tool for systematic experimentation.

Finally, a number of public agencies and Non Governmental Organizations are joining their counterparts in the private sector by providing broad-based systems thinking training to their top leaders and administrative staff. A number of simulation-based management exercises such as the production-distribution game (also known as the “beer game”) [93] and the People’s Express Flight Simulator [92] have been developed and refined over time to support such training and professional development efforts. In addition, Cavana and Clifford [11] have used GMB to develop a formal model and flight simulator to examine the policy implications of an excise tax policy on tobacco smoking.

### What Are the Arenas in Which System Dynamics Models Are Used?

The malpractice insurance vignette and the GORA example represented cases where a model was developed for a single problem within one agency. Naill [66] provides an example of how a sustained modeling capability can be installed within an agency to support a range of ongoing policy decisions (in this case the model was looking at transitional energy policies at the federal level). Barney [10] developed a class of System Dynamics simulation

models to support economic development and planning in developing nations. Wolstenholme [105] reported on efforts to support health planning within the British Health Service.

Addressing a tactical problem within a single public sector agency, while quite common, is only one of the many types of decision arenas in which System Dynamics models can be and are used to support public policy. Indeed, how a model is used in a public policy debate is largely determined by the unique characteristics of the specific decision-making arena in which the model is to be used. Some of the more common examples follow.

### Models Used to Support Inter-Agency and Inter-Governmental Collaborative Efforts

A quite different arena for the application of System Dynamics models to support the policy process occurs when an interagency or inter-governmental network of program managers must cooperate to meet a common mission. For example, Rohrbaugh [79] and Zagonel et al. [108] report a case where state and local officials from social services, labor, and health agencies combined their efforts with private and non-profit managers of day care services, health care services, and worker training and education services to plan for comprehensive reform of welfare policies in the late 1990s. These teams were seeking strategies to blend financial and program resources across a myriad of stovepipe regulations and reimbursement schemes to provide a seamless system of service to clients at the local level. To complete this task, they created a simulation model containing a wide range of system-level interactions

and tested policies in that model to find out what blend of policies might work. Policy implementation followed this model-based and simulation-supported policy design.

### Models Used to Support Expert Testimony in Courtroom Litigation

Cooper [20] presented one of the first published accounts of a System Dynamics model being used as a sort of expert witness in courtroom litigation. In the case he reported, Litton Industries was involved in a protracted lawsuit with the U. S. Navy concerning cost and time overruns in the construction of several naval warships. In a nutshell, the Navy contended that the cost overruns were due to actions taken (or not taken) by Litton Industries as primary contractor on the project and as such the Navy should not be responsible for covering cost overruns. Litton maintained that a significant number of change orders made by the Navy were the primary drivers of cost overruns and time delays. A simulation model was constructed of the ship-building process and the simulation model then built two simulated ships without any change orders. A second set of “what if” runs subsequently built the same ships except that the change orders from the Navy were included in the construction process. By running and re-running the model, the analysts were able to tease out what fractions of the cost overrun could reasonably be attributed to Litton and what fraction should be attributed to naval change orders. Managers at Litton Industries attribute their receipt of hundreds of millions dollars of court-sanctioned payments to the analysis supported by this System Dynamics simulation model. Ackermann, Eden and Williams [1] have used a similar approach involving soft systems approaches combined with a System Dynamics model in litigation over cost overruns in the channel tunnel project.

### Models Used as Part of the Legislative Process

While System Dynamics models have been actively used to support agency-level decision making, inter-agency and inter-governmental task forces and planning, and even courtroom litigation, their use in direct support of legislative processes has a more uneven track record. For example, Ford [30] reports successes in using System Dynamics modeling to support regulatory rule making in the electric power industry, and Richardson and Lamitie [73] report on how System Dynamics modeling helped redefine a legislative agenda relating to the school aid formula in the U.S. state of Connecticut. However, Andersen [4] remains more pessimistic about the ability of System Dynamics models to directly support legislative decision making, especially when the decisions involve zero-sum tradeoffs in

the allocation of resources (such as formula-driven aid involving local municipal or education formulas). This class of decisions appears to be dominated by short-term special interests. A longer-term dynamic view of such immediate resource allocation problems is less welcome. The pathway to affecting legislative decision making appears to be by working through and with public agencies, networks of providers, the courts, or even in some opinions, by directly influencing public opinion.

### Models Used to Inform the Public and Support Public Debate

In addition to using System Dynamics modeling to support decision making in the executive, judicial, and legislative branches of government (often involving Non-Governmental Organizations and private sector support), a number of System Dynamics studies appeal directly to the public. These studies intend to affect public policy by shaping public opinion in the popular press and the policy debate. In the 1960s, Jay Forrester’s *Urban Dynamics* [34] presented a System Dynamics model that looked at many of the problems facing urban America in the latter half of the 20th century. Several years later in response to an invitation from the Club of Rome, Forrester put together a study that led to the publication of *World Dynamics* [35], a highly aggregate System Dynamics model that laid out a feedback-oriented view of a hypothesized set of relationships between human activity on the planet, industrialization, and environmental degradation. Meadows et al. [61] followed on this study with a widely hailed (and critiqued) System Dynamics simulation study embodied in the best-selling book, *Limits to Growth*. Translated into over 26 languages, this volume coalesced a wide range of public opinion leading to a number of pieces of environmental reform in the decade of the 1970s. The debate engendered by that volume continues even 30 years later [63]. Donella Meadows continued in this tradition of appealing directly to public opinion through her syndicated column, *The Global Citizen*, which was nominated for the Pulitzer Prize in 1991. The column presented a System Dynamics-based view of environment matters for many years (<http://www.pcdf.org/meadows/>).

### What Are some of the Substantive Areas Where System Dynamics Has Been Applied?

The International System Dynamics Society (<http://www.systemdynamics.org>) maintains a comprehensive bibliography of over 8,000 scholarly books and articles documenting a wide variety of applications of System Dynamics modeling to applied problems in all sectors. MacDon-



ald et al. [54] have created a bibliography extracted from this larger database that summarizes some of the major areas where System Dynamics modeling has been applied to public policy. Below, we summarize some of the substantive areas where System Dynamics has been applied, giving one or two sample illustrations for each area.

### Health Care

System Dynamicists have been applying their tools to analyze health care issues at both the academic and practitioner level for many years. *The System Dynamics Review*, the official journal of the System Dynamics Society, devoted a special issue to health care in 1999 due to the importance of health care as a critical public policy issue high on the political agenda of many countries and as an area where much System Dynamics work has been performed. The extensive System Dynamics work performed in the health care area fell into three general categories: patient flow management, general health policy, and specific health problems.

The patient flow management category is exemplified by the work of Wolstenholme [106], Lane and Rosenhead [48], and Van Ackere and Smith [97]. The articles written by these authors focused on issues and policies relating to patient flows in countries where health care service is universal.

The general health policy category is rather broad in that these articles covered policy and decision making from the micro level [96] to the macro level [88]. There were also many articles that showed how the process of modeling resulted in better understanding of the problem and issues facing health care providers and policy makers [12].

The last category dealt with specific health-related problems such as the spread of AIDS [43,76], smoking [42], and malaria control [29], as well as many other health-related conditions.

### Education

The education articles touched on various topics relating to education ranging from using System Dynamics in the classroom as a student-centered teaching method to models that dealt with resource allocations in higher education. Nevertheless, many of the articles fell into five categories that could be labeled management case studies or flight simulators, teaching technology, research, teaching, and education policy.

The management case study and flight simulator articles are best exemplified by Sterman's [93] article describing the Beer Game and Graham, Morecroft et al. [41]

article on "Model Supported Case Studies for Management Education." The emphasis of these works is on the use of case studies in higher education, with the addition of games or computer simulations. This is related to the teaching technology category in that both emphasize using System Dynamics models/tools to promote learning. However, the teaching technology category of articles stresses the introduction of computer technology, specifically System Dynamics computer technology, into the classroom. Steed [90] has written an article that discusses the cognitive processes involved while using Stella to build models, while Waggoner [99] examined new technologies versus traditional teaching approaches.

In addition to teaching technology are articles that focus on teaching. The teaching category is very broad in that it encompasses teaching System Dynamics in K-12 and higher education as subject matter [37,77] as well as ways to integrate research into the higher-education classroom [71]. System Dynamics models are also used to introduce advanced mathematical concepts through simulation and visualization, rather than through equations [27,28]. In addition, lesson plans for the classroom are also part of this thread [44]. The Creative Learning Exchange (<http://www.clexchange.com>) provides a central repository of lessons and models useful for pre-college study of System Dynamics, including a selfstudy roadmap to System Dynamics principles [23].

There are also a number of articles that pertain to resource allocation [15] at the state level for K-12 schools along with articles that deal with resource-allocation decisions in higher education [32,40]. Saeed [83] and Mashayekhi [56] cover issues relating to higher education policy in developing countries.

The last education category involved research issues around education. These articles examined whether the System Dynamics methodology and simulation-based education approaches improved learning [24,47,55].

### Defense

System Dynamics modeling work around the military has focused on manpower issues, resource allocation decisions, decision making and conflict. Coyle [21] developed a System Dynamics model to examine policies and scenarios involved in sending aircraft carriers against land-based targets. Wils, Kamiya et al. [103] have modeled internal conflicts as a result of outbreaks of conflict over allocation and competition of scarce resources. The manpower articles focused on recruitment and retention policies in the armed forces and are represented in articles by Lopez and Watson [51], Andersen and Emmerichs [5],



Clark [18], Clark, McCullough et al. [19] and Cavana et al. [14]. The resource allocation category deals with issues of money and materials, as opposed to manpower, and is represented by Clark [16,17]. Decision making in military affairs from a System Dynamics perspective is represented in the article by Bakken and Gilljam [8].

## Environment

The System Dynamics applications dealing with environmental resource issues can be traced back to when the techniques developed in *Industrial Dynamics* were beginning to be applied to other fields. The publication of Forrester's *World Dynamics* in 1971 and the follow-up study *Limits to Growth* [61,62,63] used System Dynamics methodology to address the problem of continued population increases on industrial capital, food production, resource consumption and pollution. Furthermore, specific studies dealing with DDT, mercury and eutrophication of lakes were part of the Meadows et al. [59] project and appeared as stand-alone journal articles prior to being published as a collection in Meadows and Meadows [60].

The environmental applications of System Dynamics have moved on since that time. Recent work has combined environmental and climate issues with economic concerns thorough simulation experiments [25] as well as stakeholder participation in environmental issues [89]. In 2004, the *System Dynamics Review* ran a special issue dedicated to environmental issues. Cavana and Ford [13] were the editors and did a review of the System Dynamics bibliography in 2004, identifying 635 citations with the key words "environmental" or "resource." Cavana and Ford broke the 635 citations into 11 categories they identified as resources, energy, environmental, population, water, sustainable, natural resources, forest, ecology, agriculture, pollution, fish, waste, earth, climate and wildlife.

## General Public Policy

The System Dynamics field first addressed the issue of public policy with Forrester's *Urban Dynamics* [34] and the follow-up work contained in *Readings in Urban Dynamics* [58] and Alfeld and Graham's *Introduction to Urban Dynamics* [3]. The field then branched out into the previously mentioned *World Dynamics* and the follow-up studies related to that work. Moreover, the application of System Dynamics to general public policy issues began to spread into areas as diverse as drug policy [50], and the causes of patient dropout from mental health programs [49], to ongoing work by Saeed [82,84,85] on development issues in emerging economies. More recently, Saysel et al. [86] have examined water scarcity issues in

agricultural areas, Mashayekhi [57] reports on the impact on public finance of oil exports in countries that export oil and Jones et al. [46] cover the issues of sustainability of forests when no single entity has direct control.

This brief review of the literature where System Dynamics modeling has been used to address public policy issues indicates that the field is making inroads at the micro level (within government agencies) and at the macro level (between government agencies). Furthermore, work has been performed at the international level and at what could truly be termed the global level with models addressing public policy issues aimed at climate change.

## Evaluating the Effectiveness of System Dynamics Models in Supporting the Public Policy Process

System Dynamics modeling is a promising technology for policy development. But does it really work? Over the past several decades, a minor cottage industry has emerged that purports to document the successes (and a few failures) of System Dynamics models by reporting on case studies. These case studies report on successful applications and sometimes analyze weaknesses, making suggestions for improvement in future practice. Rouwette et al. [81] have compiled a meta-analysis of 107 such case-based stories.

However, as compelling as such case stories may be, case studies are a famously biased and unsystematic way to evaluate effectiveness. Presumably, failed cases will not be commonly reported in the literature. In addition, such a research approach illustrates in almost textbook fashion the full litany of both internal and external threats to validity, making such cases an interesting but unscientific compilation of war stories. Attempts to study live management teams in naturally occurring decision situations can have high external validity but almost always lack internal controls necessary to create scientifically sound insights.

Huz et al. [45] created an experimental design to test for the effectiveness of a controlled series of group-based System Dynamics cases in the public sector. They used a wide battery of pre- and post survey, interview, archival, administrative data, and qualitative observation techniques to evaluate eight carefully matched interventions. All eight interventions dealt with the integration of mental health and vocational rehabilitation services at the county level. Four of the eight interventions contained System Dynamics modeling sessions and four did not. These controlled interventions were designed to get at the impact of System Dynamics modeling on the public policy process.

Overall, Huz et al. [45] envisioned that change could take place in nine domains measured across three separate levels of analysis as illustrated in Table 1 below.

Using the battery of pre- and post test instruments, Huz found important and statistically significant results in eight of the nine domains measured. The exceptions were in domain 9 where they did not measure client outcomes, in domain 5 where “participants were not significantly more aligned in their perceptions on strategies for changes” (but were more aligned in goals), and in domain 7 where “no significant change was found with respect to structural conditions within the network” (but two other dimensions of organizational relationships did change).

In their meta-analysis of 107 case studies of System Dynamics applications, Rouwette et al. [81] coded case studies with respect to eleven classes of outcomes, sorted into individual level, group level, and organizational level. The 107 cases were dominated by for-profit examples with 65 such cases appearing in the literature followed by 21 cases in the non-profit sector, 18 cases in governmental settings, and three cases in mixed settings. While recognizing possible high levels of bias in reported cases as well as difficulties in coding across cases and a high number of missing categories, they found high percentages of positive outcomes along all 11 dimensions of analysis. For each separate dimension, they analyzed between 13 and 101 cases with the fraction of positive outcomes for each dimension ranging from a low of 83% to several dimensions where 100% of the cases reporting on a dimension found positive results. At the individual level, they coded for overall positive reactions to the work, insight

gained from the work, and some level of individual commitment to the results emerging from the study. At the group level, they coded for increased levels of communication, the emergence of shared language, and increases in consensus or mental model alignment. Organizational level outcomes included implementation of system level change. With respect to this important overall indicator they “found 84 projects focused on implementation, which suggests that in half (42) of the relevant cases changes are implemented. More than half (24) of these changes led to positive results”(see p. 20 in [81]).

Rouwette [80] followed this meta-analysis with a detailed statistical analysis of a series of System Dynamics-based interventions held mostly in governmental settings in the Netherlands. He was able to estimate a statistical model that demonstrated how System Dynamics group model building sessions moved both individuals and groups from beliefs to intentions to act, and ultimately on to behavioral change.

In sum, attempts to evaluate System Dynamics interventions in live settings continue to be plagued by methodological problems that researchers have struggled to overcome with a number of innovative designs. What is emerging from this body of study is a mixed, “good news and bad news” picture. All studies that take into account a reasonable sample of field studies show some successes and some failures. About one-quarter to one-half of the System Dynamics studies investigated showed low impact on decision making. On the other hand, roughly half of the studies have led to system-level implemented change with approximately half of the implemented studies being associated with positive measures of success.

**Public Policy, System Dynamics Applications to, Table 1**  
Domains of measurement and evaluation used to assess impact of systems-dynamics interventions (see p. 151 in [45])

Level I	Reflections of the modeling team
Domain 1	Modeling team's assessment of the intervention
Level II	Participant self-reports of the intervention
Domain 2	Participants' perceptions of the intervention
Domain 3	Shifts in participants' goal structures
Domain 4	Shifts in participants' change strategies
Domain 5	Alignment of participant mental models
Domain 6	Shifts in understanding how the system functions
Level III	Measurable system change and “bottom line” results
Domain 7	Shifts in network of agencies that support services integration
Domain 8	Changes in system-wide policies and procedures
Domain 9	Changes in outcomes for clients

### Summary: System Dynamics – A Powerful Tool to Support Public Policy

While recognizing and respecting the difficulties of scientific evaluation of System Dynamics studies in the public sector, we remain relentlessly optimistic about the method's utility as a policy design and problem-solving tool. Our glass is half (or even three-quarters) full. A method that can deliver high decision impact up to three-quarters of the time and implement results in up to half of the cases examined (and in a compressed time frame) is a dramatic improvement over alternative approaches that can struggle for months or even years without coming to closure on important policy directions.

System Dynamics-based modeling efforts are effective because they join the minds of public managers and policy makers in an emergent dialog that relies on formal modeling to integrate data, other empirical insights, and

mental models into the policy process. Policy making begins with the pre-existing mental models and policy stories that managers bring with them into the room. Policy consensus and direction emerge from a process that combines social facilitation with technical modeling and analysis. The method blends dialog with data. It begins with an emergent discussion and ends with an analytic framework that moves from “what is” baseline knowledge to informed “what if” insights about future policy directions.

In sum, we believe that a number of the process features related to building System Dynamics models to solve public policy problems contribute to their appeal for front-line managers:

- **Engagement** Key managers can be in the room as the model is evolving, and their own expertise and insights drive all aspect of the analysis.
- **Mental models** The model-building process uses the language and concepts that managers bring to the room with them, making explicit the assumptions and causal mental models managers use to make their decisions.
- **Complexity** The resulting nonlinear simulation models lead to insights about how system structure influences system behavior, revealing understandable but initially counterintuitive tendencies like policy resistance or “worse before better” behavior.
- **Alignment** The modeling process benefits from diverse, sometimes competing points of view as stakeholders can have a chance to wrestle with causal assumptions in a group context. Often these discussions realign thinking and are among the most valuable portions of the overall modeling effort.
- **Refutability** The resulting formal model yields testable propositions, enabling managers to see how well their implicit theories match available data about overall system performance.
- **Empowerment** Using the model managers can see how actions under their control can change the future of the system.

System Dynamics modeling projects merge managers’ causal and structural thinking with the available data, drawing upon expert judgment to fill in the gaps concerning possible futures. The resulting simulation models provide powerful tools to develop a shared understanding and to ground what-if thinking.

### Future Directions

While the field of System Dynamics has reached its half-century in 2007, its influence on public policy continues

to grow. Many of the problems defined by the earliest writers in the field continue to challenge us today. The growing literature base of environmental, social, and education policy is evidence of continued interest in the systems perspective. In addition, System Dynamics modeling is growing in popularity for defense analysis, computer security and infrastructure planning, and emergency management. These areas have the characteristic problems of complexity and uncertainty that require the integration of multiple perspectives and tacit knowledge that this method supports. Researchers and practitioners will continue to be attracted to the open nature of System Dynamics models as a vehicle for consensus and experimentation.

We anticipate that the tool base for developing and distributing System Dynamics models and insights will also grow. Graphical and multimedia-based simulations are growing in popularity, making it possible to build clearer models and disseminate insights easily. In addition, the development of materials for school-age learners to consider a systems perspective to social problems gives us optimism for the future of the field, as well as for future policy.

### Bibliography

#### Primary Literature

1. Ackermann F, Eden C, Williams T (1997) Modeling for Litigation: Mixing Qualitative and Quantitative Approaches. *Interfaces* 27(2):48–65
2. Akkermans H, Vennix J (1997) Clients’ Opinions on Group Model-Building: An Exploratory Study. *Syst Dyn Rev* 13(1):3–31
3. Alfeld L, Graham A (1976) *Introduction to Urban Dynamics*. Wright-Allen Press, Cambridge
4. Andersen D (1990) Analyzing Who Gains and Who Loses: The Case of School Finance Reform in New York State. *Syst Dyn Rev* 6(1):21–43
5. Andersen D, Emmerichs R (1982) Analyzing US Military Retirement Policies. *Simulation* 39(5):151–158
6. Andersen D, Richardson GP (1997) Scripts For Group Model Building. *Syst Dyn Rev* 13(2):107–129
7. Andersen D, Bryson J, Richardson GP, Ackermann F, Eden C, Finn C (2006) Integrating Modes of Systems Thinking into Strategic Planning Education and Practice: The Thinking Persons’ Institute Approach. *J Public Aff Educ* 12(3):265–293
8. Bakken B, Gilljam M (2003) Dynamic Intuition in Military Command and Control: Why it is Important, and How It Should be Developed. *Cogn Technol Work* (5):197–205
9. Barlas Y (2002) System Dynamics: Systemic Feedback Modeling for Policy Analysis. In: *Knowledge for Sustainable Development, an Insight into the Encyclopedia of Life Support Systems*, vol 1. UNESCO-EOLSS, Oxford, pp 1131–1175
10. Barney G (1982) *The Global 2000 Report to the President: Entering the Twenty-First Century*. Penguin, New York
11. Cavana RY, Clifford L (2006) Demonstrating the utility for system dynamics for public policy analysis in New Zealand:

- the case for excise tax policy on tobacco. *Syst Dyn Rev* 22(4):321–348
12. Cavana RY, Davies P et al (1999) Drivers of Quality in Health Services: Different Worldviews of Clinicians and Policy Managers Revealed. *Syst Dyn Rev* 15(3):331–340
  13. Cavana RY, Ford A (2004) Environmental and Resource Systems: Editor's Introduction. *Syst Dyn Rev* 20(2):89–98
  14. Cavana RY, Boyd D, Taylor R (2007) A Systems Thinking Study of Retention and Recruitment Issues for the New Zealand Army Electronic Technician Trade Group. *Syst Res Behav Sci* 24(2):201–216
  15. Chen F, Andersen D et al (1981) A Preliminary System Dynamics Model of the Allocation of State Aid to Education. *Dynamica* 7(1):2–13
  16. Clark R (1981) Readiness as a Residual of Resource Allocation Decisions. *Def Manag J* 1:20–24
  17. Clark R (1987) Defense Budget Instability and Weapon System Acquisition. *Public Budg Financ* 7(2):24–36
  18. Clark R (1993) The Dynamics of US Force Reduction and Reconstitution. *Def Anal* 9(1):51–68
  19. Clark T, McCullough B et al (1980) A Conceptual Model of the Effects of Department of Defense Realignments. *Behav Sci* 25(2):149–160
  20. Cooper K (1980) Naval Ship Production: A Claim Settled and Framework Built. *Interfaces* 10(6):20
  21. Coyle RG (1992) A System Dynamics Model of Aircraft Carrier Survivability. *Syst Dyn Rev* 8(3):193–213
  22. Coyle RG (1996) System Dynamics Modelling: A Practical Approach. Chapman and Hall, London
  23. Creative Learning Exchange (2000) Road Maps: A Guide to Learning System Dynamics. Available at <http://sysdyn.clexchange.org/road-maps/home.html>
  24. Davidsen P (1996) Educational Features of the System Dynamics Approach to Modelling and Learning. *J Struct Learn* 12(4):269–290
  25. Fiddaman TS (2002) Exploring Policy Options with a Behavioral Climate-Economy Model. *Syst Dyn Rev* 18(2):243–267
  26. Finely B (2006) Migrant Cases Burden System: Rise in Deportations Floods Detention Centers, Courts. *Denver Post*, 10/2/06: [http://www.denverpost.com/immigration/ci\\_4428563](http://www.denverpost.com/immigration/ci_4428563)
  27. Fisher D (2001) Lessons in Mathematics: A Dynamic Approach. iSee Systems, Lebanon
  28. Fisher D (2004) Modeling Dynamic Systems: Lessons for a First Course. iSee Systems, Lebanon
  29. Flessa S (1999) Decision Support for Malaria-Control Programmes – a System Dynamics Model. *Health Care Manag Sci* 2(3):181–91
  30. Ford A (1997) System Dynamics and the Electric Power Industry. *Syst Dyn Rev* 13(1):57–85
  31. Ford A (1999) Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems. Island Press, Washington, DC
  32. Forsyth B, Hirsch G et al (1976) Projecting a Teaching Hospital's Future Utilization: A Dynamic Simulation Approach. *J Med Educ* 51(11):937–9
  33. Forrester J (1961) *Industrial Dynamics*. Pegasus Communications, Cambridge
  34. Forrester J (1969) *Urban Dynamics*. Pegasus Communications, Waltham
  35. Forrester J (1971) *World Dynamics*. Pegasus Communications, Waltham
  36. Forrester J (1980) Information Sources for Modeling the National Economy. *J Am Stat Assoc* 75(371):555–566
  37. Forrester J (1993) System Dynamics as an Organizing Framework for Pre-College Education. *Syst Dyn Rev* 9(2):183–194
  38. Forrester J (1994) Policies, Decisions, and Information Sources for Modeling. Modeling for Learning Organizations. In: Morecroft J, Sterman J (eds) Productivity Press. Portland, OR, pp 51–84
  39. Forrester J, Senge P (1980) Tests for Building Confidence in System Dynamics Models. In: Legasto Jr. AA et al (eds) *System Dynamics*. North-Holland, New York, 14, pp 209–228
  40. Galbraith P (1989) Mathematics Education and the Future: A Long Wave View of Change. *Learn Math* 8(3):27–33
  41. Graham A, Morecroft J et al (1992) Model Supported Case Studies for Management Education. *Eur J Oper Res* 59(1):151–166
  42. Homer J, Roberts E et al (1982) A Systems View of the Smoking Problem. *Int J Biomed Comput* 13:69–86
  43. Homer J, St Clair C (1991) A Model of HIV Transmission Through Needle Sharing. A Model Useful in Analyzing Public Policies, Such as a Needle Cleaning Campaign. *Interfaces* 21(3):26–29
  44. Hopkins P (1992) Simulating Hamlet in the Classroom. *Syst Dyn Rev* 8(1):91–100
  45. Huz S, Andersen D, Richardson GP, Boothroyd R (1997) A Framework for Evaluating Systems Thinking Interventions: An Experimental Approach to Mental Health System Change. *Syst Dyn Rev* 13(2):149–169
  46. Jones A, Seville D et al (2002) Resource Sustainability in Commodity Systems: The Sawmill Industry in the Northern Forest. *Syst Dyn Rev* 18(2):171–204
  47. Keys B, Wolfe J (1996) The Role of Management Games and Simulations in Education Research. *J Manag* 16(2):307–336
  48. Lane DC, Monefeldt C, Rosenhead JV (1998) Emergency – But No Accident – A System Dynamics Study of an accident and emergency department. *OR Insight* 11(4):2–10
  49. Levin G, Roberts E (eds) (1976) *The Dynamics of Human Service Delivery*. Ballinger, Cambridge
  50. Levin G, Hirsch G, Roberts E (1975) *The Persistent Poppy: A Computer Aided Search for Heroin Policy*. Ballinger, Cambridge
  51. Lopez T, Watson J Jr (1979) A System Dynamics Simulation Model of the U. S. Marine Corps Manpower System. *Dynamica* 5(2):57–78
  52. Luna-Reyes L, Martinez-Moyano I, Pardo T, Creswell A, Richardson GP, Andersen D (2007) Anatomy of a Group Model Building Intervention: Building Dynamic Theory from Case Study Research. *Syst Dyn Rev* 22(4):291–320
  53. Maani KE, Cavana RY (2007) *Systems Thinking, System Dynamics: Managing Change and Complexity*. Pearson Education (NZ) Ltd, Auckland
  54. MacDonald R et al (2007) System Dynamics Public Policy Literature. *Syst Dyn Soc*, [http://www.systemdynamics.org/short\\_bibliography.htm](http://www.systemdynamics.org/short_bibliography.htm)
  55. Mandinach E, Cline H (1993) Systems, Science, and Schools. *Syst Dyn Rev* 9(2):195–206
  56. Mashayekhi A (1977) Economic Planning and Growth of Education in Developing Countries. *Simulation* 29(6):189–197
  57. Mashayekhi A (1998) Public Finance, Oil Revenue Expenditure and Economic Performance: A Comparative Study of Four Countries. *Syst Dyn Rev* 14(2–3):189–219



58. Mass N (ed) (1974) *Readings in Urban Dynamics*. Wright-Allen Press, Cambridge
59. Meadows D, Beherens W III, Meadows D, Nail R, Randers J, Zahn E (ed) (1974) *Dynamics of Growth in a Finite World*. Pegasus Communications, Waltham
60. Meadows D, Meadows D (1977) *Towards Global Equilibrium: Collected Papers*. MIT Press, Cambridge
61. Meadows D, Meadows D, Randers J (1972) *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Universe Books, New York
62. Meadows D, Meadows D, Randers J (1992) *Beyond the Limits*. Chelsea Green Publishing Company, Post Mills
63. Meadows D, Randers J, Meadows D (2004) *Limits to Growth: The 30-Year Update*. Chelsea Green Publishing Company, White River Junction
64. Miyakawa T (ed) (1999) *The Science of Public Policy: Essential Readings in Policy Sciences 1*. Routledge, London
65. Morecroft J (2007) *Strategic Modelling and Business Dynamics: A Feedback Systems Approach*. Wiley, West Sussex
66. Naill R (1977) *Managing the Energy Transition*. Ballinger Publishing Co, Cambridge
67. Otto P, Struben J (2004) Gloucester Fishery: Insights from a Group Modeling Intervention. *Syst Dyn Rev* 20(4):287–312
68. Reagan-Cirincione P, Shuman S, Richardson GP, Dorf S (1991) Decision modeling: Tools for Strategic Thinking. *Interfaces* 21(6):52–65
69. Richardson GP (1991) System Dynamics: Simulation for Policy Analysis from a Feedback Perspective. In: Fishwick P, Luker P (eds) *Qualitative Simulation Modeling and Analysis*. Springer, New York
70. Richardson GP (1996) System Dynamics. In: Gass S, Harris C (eds) *Encyclopedia of Operations Research and Management Science*. Kluwer Academic Publishers, Norwell
71. Richardson G, Andersen D (1979) Teaching for Research in System Dynamics. *Dynamica* 5(3)
72. Richardson G, Andersen D (1995) Teamwork in Group Model Building. *Syst Dyn Rev* 11(2):113–137
73. Richardson G, Lamitie R (1989) Improving Connecticut School Aid: A Case Study with Model-Based Policy Analysis. *J Educ Financ* 15(2):169–188
74. Richardson G, Pugh J (1981) *Introduction to System Dynamics Modeling*. Pegasus Communications, Waltham
75. Richmond B (1997) The Strategic Forum: Aligning Objectives Strategy and Process. *Syst Dyn Rev* 13(2):131–148
76. Roberts C, Dangerfield B (1990) Modelling the Epidemiological Consequences of HIV Infection and AIDS: a Contribution from Operational Research. *J Oper Res Soc* 41(4):273–289
77. Roberts N (1983) An Introductory Curriculum in System Dynamics. *Dynamica* 9(1):40–42
78. Roberts N, Andersen DF, Deal RM, Grant MS, Schaffer WA (1983) *Introduction to Computer Simulation: a System Dynamics Modeling Approach*. Addison Wesley, Reading
79. Rohrbaugh J (2000) The Use of System Dynamics in Decision Conferencing: Implementing Welfare Reform in New York State. In: Garson G (ed) *Handbook of Public Information Systems*. Marcel Dekker, New York, pp 521–533
80. Rouwette E (2003) *Group Model Building as Mutual Persuasion*. Wolf Legal Publishers, Nijmegen
81. Rouwette E, Vennix J, Van Mullekom T (2002) Group Model Building Effectiveness: a Review of Assessment Studies. *Syst Dyn Rev* 18(1):5–45
82. Saeed K (1994) *Development Planning and Policy Design: A System Dynamics Approach*. Ashgate, Aldershot
83. Saeed K (1996) The Dynamics of Collegial Systems in the Developing Countries. *High Educ Policy* 9(1):75–86
84. Saeed K (1998) *Towards Sustainable Development, 2nd Edition: Essays on System Analysis of National Policy*. Ashgate, Aldershot
85. Saeed K (2003) *Articulating Developmental Problems for Policy Intervention: A System Dynamics Modeling Approach*. *Simul Gaming* 34(3):409–436
86. Saysel A, Barlas Y et al (2002) Environmental Sustainability in an Agricultural Development Project: A System Dynamics Approach. *J Environ Manag* 64(3):247–260
87. Senge P (1990) *The Fifth Discipline: the Art and Practice of the Learning Organization*. Doubleday/Currency, New York
88. Senge P, Asay D (1988) Rethinking the Healthcare System. *Healthc Reform J* 31(3):32–34, 44–45 5
89. Stave K (2002) Using SD to Improve Public Participation in Environment Decisions. *Syst Dyn Rev* 18(2):139–167
90. Steed M (1992) Stella, a Simulation Construction Kit: Cognitive Process and Educational Implications. *J Comput Math Sci Teach* 11(1):39–52
91. Stenberg L (1980) A Modeling Procedure for the Public Policy Scene. In: Randers J (ed) *Elements of the System Dynamics Method*. Pegasus Communications, Waltham, pp 257–288
92. Sterman J (1988) *People Express Management Flight Simulator: Simulation Game, Briefing Book, and Simulator Guide*. <http://web.mit.edu/jsterman/www/SDG/MFS/PE.html>
93. Sterman J (1992) Teaching Takes Off: Flight Simulators for Management Education. *OR/MS Today* (October):40–44
94. Sterman J (1996) A Skeptic's Guide to Computer Models. In: Richardson GP (ed) *Modelling for Management*. Dartmouth Publishing Company, Aldershot
95. Sterman J (2000) *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill, Boston
96. Taylor K, Lane D (1998) *Simulation Applied to Health Services: Opportunities for Applying the System Dynamics Approach*. *J Health Serv Res Policy* 3(4):226–232
97. van Ackere A, Smith P (1999) Towards a Macro Model of National Health Service Waiting Lists. *Syst Dyn Rev* 15(3):225–253
98. Vennix J (1996) *Group model building: Facilitating team learning using system dynamics*. Wiley, Chichester
99. Waggoner M (1984) The New Technologies versus the Lecture Tradition in Higher Education: Is Change Possible? *Educ Technol* 24(3):7–13
100. Warren K (1999) Dynamics of Strategy. *Bus Strateg Rev* 10(3):1–16
101. Warren K (2002) *Competitive Strategy Dynamics*. Wiley, Chichester
102. Warren K (2005) Improving Strategic Management with the Fundamental Principles of System Dynamics. *Syst Dyn Rev* 21(4):329–350
103. Wils A, Kamiya M et al (1998) Threats to Sustainability: Simulating Conflict Within and Between Nations. *Syst Dyn Rev* 14(2–3):129–162
104. Wolstenholme E (1992) The Definition and Application of a Stepwise Approach to Model Conceptualization and Analysis. *Eur J Oper Res* 59(1):123–136
105. Wolstenholme E (1993) A Case Study in Community Care Using Systems Thinking. *J Oper Res Soc* 44(9):925–934



106. Wolstenholme E (1999) A Patient Flow Perspective of UK Health Services: Exploring the Case for New Intermediate Care Initiatives. *Syst Dyn Rev* 15(3):253–273
  107. Wolstenholme E, Coyle RG (1983) The Development of System Dynamics as a Methodology for System Description and Qualitative Analysis. *J Oper Res Soc* 34(7):569–581
  108. Zagonel A, Andersen D, Richardson GP, Rohrbaugh J (2004) Using Simulation Models to Address “What If” Questions about Welfare Reform. *J Policy Anal Manag* 23(4):890–901
- Books and Reviews**
- Coyle RG (1998) *The Practice of System Dynamics: Milestones, Lessons and Ideas from 30 Years of Experience*. *Syst Dyn Rev* 14(4):343–365
- Forrester J (1961) *Industrial Dynamics*. Pegasus Communications, Waltham
- Maani KE, Cavana RY (2007) *Systems Thinking, System Dynamics: Managing Change and Complexity*. Pearson Education (NZ) Ltd, Auckland
- MacDonald R (1998) Reducing Traffic Safety Deaths: A System Dynamics Perspective. In: 16th International Conference of the System Dynamics Society, Quebec '98, System Dynamics Society Quebec City
- Morecroft JDW, Sterman JD (eds) (1994) *Modeling for Learning Organizations*, System Dynamics Series. Productivity Press, Portland
- Wolstenholme EF (1990) *System Enquiry: A System Dynamics Approach*. Wiley, Chichester