

## N

## Nanocomputers

FERDINAND PEPPER

National Institute of Information  
and Communications Technology, Kobe, Japan

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Wires and Devices  
Nanofabrication Techniques and Architectures  
Heat Dissipation  
Fault-Tolerance  
Cellular Automaton-Based Architectures  
Crossbar Array-Based Architectures  
Neural Network-Based Architectures  
Future Directions  
Bibliography

### Glossary

**Adiabatic switching** Switching with asymptotically zero speed with the aim of reducing power consumption in a circuit.

**Asynchronous circuit** Circuit that is designed to work in the absence of a clock.

**Babbage engine** Mechanical calculator built by Charles Babbage in the early 19th century.

**Bottom-up fabrication** Fabrication method employing the natural ability of physical structures (including atoms and molecules) to organize themselves into desired structures.

**Brownian motion** Random movement of micrometer-sized particles due to collisions with molecules. The term is also used to indicate random movement of smaller-sized particles, or the mathematical model of such movements.

**Carbon nanotube** Nanometer-scale tube consisting of a graphite sheet rolled up into a seamless cylinder. Carbon nanotubes intended for nanoelectronic applications are mainly single-walled.

**Cellular automaton** Discrete regular array of cells, each of which is in one of a finite number of states. A cell is updated in discrete time steps according to a transition rule that takes the states of the cell and its direct neighbors and determines the next state of the cell. If all cells are updated at the same time, the model is called *synchronous*, otherwise *asynchronous*.

### Complementary metal-oxide-semiconductor (CMOS)

Currently dominant technology used to implement digital logic circuits as well as a wide variety of analog circuits such as image sensors and data convertors.

**Computational universality** Ability of a computing system to compute every function in a certain class of systems.

**Computer architecture** Functional and structural design and operational specification of a computer system.

**Construction universality** Ability of a logic system to construct every arbitrary logic structure in a certain class of systems.

**Crossbar array** Array consisting of two layers of wires, whereby wires within one layer are parallel to each other but perpendicular to the wires in the other layer. The wires are connected to each other at their cross-points through devices.

**Coulomb blockade** Increased resistance at certain voltages to the flow of electrons in a *tunnel-junction*, which is a thin insulating barrier between two conducting electrodes.

**Defect-tolerant** Ability of a system to remain relatively unaffected by the occurrence of permanent defects.

**Delay-insensitive circuit** Asynchronous circuit in which the outcomes of operations are unaffected by delays in wires and functional elements.

**Device** Functional element in a circuit that takes input signals and produces output signals. A transistor is a well-known example of a device.

- Dielectric material** Medium that is a poor conductor of electricity, but an effective supporter of electric fields. A dielectric with a high dielectric constant  $\kappa$  is the preferred material used as a gate-dielectric in a transistor, since it allows for a thicker insulating layer between gate and channel given a certain gate capacitance. A low- $\kappa$  dielectric is the preferred material used for insulating layers between wires, since it allows for smaller wire pitches.
- Entropy** Measure of uncertainty associated with a certain information-theoretic or physical variable.
- Error correcting code (ECC)** Encoding that adds redundancy to information to increase the ability to correct or detect errors caused by noise or other impairments.
- Fault-tolerant** Ability of a system to remain relatively unaffected by the occurrence of errors.
- Field-effect transistor (FET)** Transistor in which the conductivity of the channel depends on the electric field controlled by the transistor's gate.
- Fine-grained parallelism** Scheme for the subdivision of tasks in a large number of small subtasks that can be simultaneously executed by a large number of simple information processing elements.
- Finite automaton** Logical scheme consisting of a finite number of states, including a start state and an accept state, as well as a specification of state changes under the influence of inputs.
- Field-programmable gate array (FPGA)** Type of chip containing logic components and programmable wires that can be programmed to perform a wide variety of combinational logic functions.
- Gate** A logic gate is a digital device that carries out a Boolean bit-operation, such as the AND, OR, NOT, etc. A transistor gate is the part of the transistor that controls the conductivity of the channel between the transistor's source and drain.
- Hamiltonian** Property corresponding to the total energy of a system's state that is determined by some sequence of physical operations. A Hamiltonian is mathematically expressed by a unitarian operator  $H$ .
- Heisenberg uncertainty principle** Relationship in quantum mechanics, giving a lower bound on the product of the uncertainties of two physical observables. These two observables may be position and momentum, or, alternatively, energy and time.
- Lithography** Microfabrication process in which a pattern is transferred to a photosensitive material to selectively remove parts of a thin substrate.
- Majority gate** A logic gate in which the value of the output bit is set to the logic value occurring in the majority of the input bits. A majority gate usually has an odd number of input bits.
- Markov chain** Discrete-time stochastic process of which the next state solely depends on the present state and not directly on previous states. In other words, the process is *memoryless*.
- Molecular electronics** Electronics in which the components (wires and devices) are realized in terms of molecules. These molecules are usually organic.
- Moore's law** Trend along which integration density of microelectronics has developed since the 1960's, the time Gordon E. Moore first observed this trend. Integration density according to this trend increases exponentially, doubling approximately every two years.
- MOSFETS** FETs implemented by CMOS technology.
- Neural network** Mathematical model based on the biological nervous system, consisting of *neurons* that receive (usually) analog values from each other through weighted interconnections. Learning in a neural network takes place through updating the weights based on the values of the neurons and the values of the input signals.
- Neuron** Biological neuron is a cell in the nervous system that processes and transmits information. The central part of a neuron is its *soma* (cell body), and it has an extension called *axon* to transmit information to other neurons via axon terminals called *synapses*. This information is received by a neuron through its *dendrites*. A neuron in an artificial neural network is modeled after a biological neuron. It receives inputs via weighted interconnections, which model the strengths of synapses.
- NP-complete** Class of decision problems for which no polynomial-time (expressed in terms of the input size) algorithms are known. Any member of the wider class of NP (Non-deterministic Polynomial time) problems can be transformed in an NP-complete problem, in which the time overhead of the translation is at most a polynomial factor.
- Parallel** A computation is parallel if it is divided in smaller computations that can be executed simultaneously.
- Perceptron** A type of neural network model invented in 1957 by Frank Rosenblatt that is used for classification. Though originally this model consisted only of a layer of input neurons and a layer of output neurons, modern uses of the term includes the possible presence of one or more hidden layers of neurons. The perceptron is a *feedforward* neural network, which means that information flows in one direction, from the input to the

output, so there are no backward connections between neurons.

**Pitch** Term used in integrated circuits to denote distance between elements, such as between cells in RAM memory or between the centers of two wires. Commonly used is the term *half-pitch*, indicating half this distance.

**Processor** The computing part of a computer also called Central Processing Unit (CPU).

**Quantum dot cellular automaton** Cellular automaton in which cells based on quantum dots containing electrons interact with each other through electrostatic forces.

**Ratchet** Device that allows a process (such as the movement of particles) to take place in only one direction.

**Repeater** Logic device placed on a wire to reproduce signals input to it. Usually implemented in terms of NOT-gates, repeaters speed up the propagation of signals along wires in highly integrated microelectronics.

**Resonant clock** Timing mechanism on synchronous chips that employs the resonance of oscillators to achieve sharply reduced power consumption and increased preciseness in timing.

**Resonant tunneling device** Device using quantum effects to allow very efficient transmission of electrons through a double barrier tunneling structure.

**Reversible computation** Computation of a function that is one-to-one.

**Scanning tunneling microscope (STM)** Type of electron microscope to view surfaces at the atomic level with resolutions of up to 0.1 nm lateral and 0.01 nm in depth. The STM employs a tip from which electrons tunnel to the surface, whereby the tunnel current depends on the distance between the tip and the surface as an exponential function. The STM can also be used to manipulate individual atoms and molecules.

**Single electron tunneling device** Device based on the tunneling of individual electrons through one or more tunneling barriers, which are thin insulating layers between electrodes.

**Spintronic device** Device based on the magnetic spin states of electrons.

**Superposition of states** Linear combination of states in a quantum system describing a situation in which a physical observable possesses two or more values simultaneously.

**Synapse** Receptor on a neuron's axon that connects to a dendrite of a neuron to transmit information.

**Top-down fabrication** Fabrication method in which structures are formed under the control of a master plan. Optical lithography is the usual top-down method for fabricating microelectronic chips.

**Tunneling** Quantum mechanical phenomenon in which a particle passes through an energy barrier that would, given the particle's kinetic energy, be too high for it to pass by classical physical laws.

**Tunneling phase logic** Logic that uses the phases of waves to conduct logic operations.

**Turing machine** Abstract logic model consisting of a tape, a reading and writing head, and a finite automaton to control the head. This simple machine model is used to study computation and the relations between computational models.

**Very large scale integration (VLSI)** Microelectronics technology with millions of devices on a chip.

**Voltage encoding** Encoding of the value of a signal by the level of a voltage. This is the most commonly used method to encode signals in electronics. Opposite of *charge encoding*, in which the value of a signal is determined by the presence of a small number of elementary electrical charges.

**von Neumann neighborhood** Set of cells in a cellular automaton that neighbors orthogonally to a cell. The von Neumann neighborhood of a cell in a 2-dimensional cellular automaton consists of the cell's northern, eastern, southern, and western neighboring cells. Often the cell itself is also included in the definition of neighborhood.

## Definition of the Subject

Nanocomputers are (not-yet-realized) computers that will be based on technology employing devices and wires with feature sizes in the order of a few nanometers ( $10^{-9}$  m). If the increase in integration density of microelectronics according to Moore's law [143] continues at the same pace as it has for almost 40 years, such computers will be around in a few decades. Computational power and speed of nanocomputers will likely dwarf those of most contemporary computers if trends from the past continue. It is anticipated that silicon-based CMOS technology can be extended to up to the year 2015; beyond that, major scientific and technological breakthroughs will be required, according to the International Technology Roadmap for Semiconductors (ITRS) [94]. Many of these breakthroughs will take place on the physical level, via the development of new devices, but progress on an architectural and algorithmic level will also be indispensable. Among the issues that need to be addressed in this context are the following:

- How to build circuitry with feature sizes well below what can be potentially realized by extensions of traditional manufacturing technologies, like those based on optical lithography, even if taking into account major

advances in technology? Will manufacturing based on molecular self-assembly and self-organization be able to deliver such a feat? What will the impact of the adoption of such bottom-up manufacturing have on the architectures of computers built by it? Will it require structures to be highly regular or random, and how will this affect the way such computers are configured and programmed?

- How to cope with heat dissipation, which will be especially debilitating at the extremely high integration densities expected in nanocomputers? Cooling techniques seem to be insufficient in this context, leaving only the reduction of heat dissipation as an option, which will inevitably involve the reduction of energy consumption. Will it be sufficient to this end to develop new devices, like carbon nanotube-based devices, resonant tunneling devices, single-electron tunneling devices, spintronic devices, molecular devices, etc., or will other innovations be required? Will it be necessary to use asynchronous circuits, i.e., circuits that work without a clock? Alternatively, will it be necessary and possible to use reversible and adiabatic computing schemes (► [Reversible Computing](#))? And how will these schemes affect a nanocomputer's architecture?
- How to cope with the unprecedented rates of errors and manufacturing defects that are expected in nanocomputers? In case of manufacturing defects, will it be possible to reconfigure a nanocomputer around such defects, and if so, how to detect the locations of such defects and how to reconfigure circuitry around them? In case of transient errors, which are associated with noise and fluctuations, and which are expected to occur at rates far beyond those experienced in CMOS technology, how to resolve such errors?

Some of above issues have attracted interest in the past, but, ironically, the success of CMOS technology has been a delaying factor in the initiation of systematic studies into them.

## Introduction

In his 1959 lecture entitled *There's plenty of Room at the Bottom* [60], Feynman predicted a future in which a single bit would require a mere 100 atoms to be stored. The same lecture pointed out the possibility of computers with devices miniaturized to similar scales. Nowadays such machines are called *nanocomputers*. Feynman's proposal, though visionary, was not made in a vacuum. Molecular conduction, which is of paramount importance to molecular electronics, was investigated by Mulliken and Szent-Gyorgi in preceding decades [89]. Fur-

thermore, as early as 1956 Arthur von Hippel advanced his view toward the engineering of materials from atoms and molecules [200], based on a firm theoretical understanding of material properties. These ideas made their way into the electronics industry through Westinghouse, which started a Molecular Systems Engineering program in 1957 [33]. The first conference on Molecular Electronics was held in 1958, jointly organized by the Air Research and Development Command, as well as by the National Security Industrial Association, and attracting approximately 300 participants [33]. The interest in military circles for this new field stemmed especially from the need to reduce size, weight, and cost of devices, as well as to increase reliability of electronics, which at the time was far worse than today. Westinghouse's endeavor in molecular electronics received funding from 1959 to 1962 from the US Air Force, but it was discontinued due to a lack of concrete results, especially with respect to manufacturing issues, though it appears to have had a positive spin-off on clean room techniques and scanning electron microscopy [33]. The effort disappeared silently from public view in successive years (see p. 181 in [30]). In retrospect, the idea was far ahead of its time.

A decade later, in the mid 1970's, molecular electronics received renewed attention, when Aviram pursued his radical vision of devices based on individual molecules together with Ratner [6]. Around the same time, Carter began work on molecular wires, switches, molecular logic elements, and molecular computers [26,27,28], thereby becoming one of the pioneers of nanocomputer architecture.

The 1980's saw an intense interest in the use of physics for the implementation of computation [12,19,65], with special focus on cellular automaton-like architectures. Drexler, around the same time, started to popularize nanotechnology and computers built by it to a general audience [55,56].

In the 1990's, interest widened to a variety of architectures and methods with potential for nanocomputing. Reliability issues of nanocomputers also started to attract attention. Representative of these efforts is the Teramac computer [84].

This trend appears to continue in the first decade of the 21st century, with unconventional computing models finding an increasing audience.

Though nanocomputers have always been overshadowed by their silicon-based counterparts, they receive attention due to CMOS technology being thought of in danger of running into serious physical limitations. Some predictions for limitations over the years are listed in Table 1 (see [97]). Though most of the expected limitations in the past have been overcome through a reexamination



**Nanocomputers, Table 1**

Past predicted limitations on gate lengths for downscaling of MOSFETs. The listed values are of gate lengths, which are smaller than the generation size listed in the International Technology Roadmap for Semiconductors (ITRS) [94]. For example, a 30 nm gate length corresponds to the 100 nm technology generation. The term *red brick wall* is named after the color-coded tables in the ITRS report, and indicates a limit beyond which no technology solutions are known (at the time)

Period	Expected limit	Cause
Late 1970's	1000 nm	Short-channel effects, lithography limitations
Early 1980's	500 nm	Source / Drain resistance
1980	250 nm	Tunneling leakage, dopant fluctuation [136]
Late 1980's	100 nm	Red brick wall: various
1999	30 nm	Lower limit to supply voltage [43]
2004	50 nm	Red brick wall: various
2004	10 nm	Fundamental

of the assumptions underlying them, coupled with human ingenuity, this may become difficult as the limitations increasingly assume a fundamental physical nature. Important fundamental physical limits include the following [127,138]:

**Thermal limit** refers to the energy required for a binary switching operation, which should at least exceed the energy of thermal fluctuations, less an error occurs. This amounts to an energy of  $(\ln 2)kT$ , whereby  $k = 1.38 \cdot 10^{-23}$  J/K is the Boltzmann constant and  $T$  the temperature in Kelvin. The probability  $P$  of a thermal fluctuation of energy  $E_s$  within a response time  $\tau$  of the circuit follows the Boltzmann relation [136]  $P = \exp[-E_s/kT]$ , which implies a rate of  $(1/\tau) \exp[-E_s/kT]$  failures per second per device. The probability due to error can thus be reduced by increasing the signal energy. To be on the safe side, a limit of  $100kT$  per switching operation is usually assumed. The limit associated with this minimum level of reliability is still four orders of magnitude below the switching energy for 100 nm CMOS technology.

**Quantum limit** relates the signal switching energy transfer  $\Delta E$  to the transition time  $t$  via the Heisenberg uncertainty principle,  $\Delta E \cdot t \geq h$ , where  $h \approx 6.6260 \times 10^{-34}$  is Planck's constant. Resulting from the wave nature of the electron and the associated uncertainty in the energy-time (or the position-momentum) relations, this limit imposes an upper bound on device switching speed given a certain average switching energy. CMOS circuits operate far above the quantum limit.

**Speed-of-light limit** imposes an upper bound on the distance traveled by a signal in one clock cycle. Given that electrical signals travel half the speed of light

( $c \approx 3 \cdot 10^8$  m/s) in typical materials, a distance of merely a few centimeters can be covered within one clock cycle on a 10 GHz chip.

An extensive analysis of limits, ranging from a fundamental level, materials, devices, circuits, up to systems, is given in [138].

Nanocomputers come in various flavors and are distinguished according to the underlying mechanisms involved. Montemerlo et al. [142] distinguish four types:

**Electronic Nanocomputers** The natural successors of electronic digital computers, these computers represent and process information through the manipulation of electrons. Though conventional transistor technology will be prevalent in the coming decade, alternatives that use tunneling of electrons or their spins have received serious attention, because of their low power potential.

**Mechanical Nanocomputers** These computers are miniaturized equivalents of Babbage engines that conduct their operations through moving molecular-scale rods and rotating molecular-scale wheels, as envisioned in [55,56]. When combined with electrical signals, switches in such computers have excellent On/Off ratios. The high mass of atoms, as compared to electrons, however, gives these computers a speed disadvantage to their electronic counterparts.

**Chemical Nanocomputers** Based on processes involving making and breaking chemical bonds to store and process information, these computers have extremely small devices, but they also will likely be slow. Most efforts in this framework find their roots in biochemistry. DNA-computers [3] (► [DNA Computing](#)) have attracted much interest in this context, though their much-touted use for solving NP-complete problems appears to run into practical limits [81].

**Quantum Nanocomputers** Quantum nanocomputers (► [Quantum Computing](#)) aim to exploit quantum effects such as superpositions of states to achieve vastly improved performance for certain algorithms (see [176] for an overview of recent progress in quantum algorithms). Though quantum effects increasingly play a role as feature sizes decrease, they have mostly be considered of significance in the context of devices, rather than in a quantum computational framework. Quantum nanocomputers seem not to be within the reach of practical applications before the year 2020 [8].

This chapter focuses mostly on electronic nanocomputers and mechanical nano-computers, these being the most likely successors of current computers. It is organized as follows. After an overview of devices and wires for nanoscale implementations in Sect. “[Wires and Devices](#)”, nanofabrication techniques are discussed in Sect. “[Nanofabrication Techniques and Architectures](#)”, because they have a direct impact on the architectures of nanocomputers. Top-down techniques, such as optical lithography, face serious limitations in this context. Unsurprisingly there is an increasing interest in bottom-up techniques, which can fabricate structures varying from highly regular on one hand to highly random on the other hand. Heat dissipation, and especially strategies to minimize it, will be the topic of Sect. “[Heat Dissipation](#)”. Asynchronous timing and reversible logic receive particular attention, but computation schemes that aim to cope and even to exploit noise are also discussed. Section “[Fault-Tolerance](#)” discusses fault-tolerance, distinguishing faults occurring during operation – which can be corrected by employing redundancy of some resource in the architecture – and faults of a permanent character, caused during manufacturing – which tend to lend themselves more to reconfiguration techniques. Following these basic topics are three sections describing architectures with particular promise for nanocomputers, i.e., Cellular Automata in Sect. “[Cellular Automaton-Based Architectures](#)”, Crossbar Arrays in Sect. “[Crossbar Array-Based Architectures](#)”, and Neural Networks in Sect. “[Neural Network-Based Architectures](#)”. This chapter finishes with a discussion about future directions.

## Wires and Devices

Three decades of development has left the basic concepts and structures of wires and devices implemented by CMOS relatively unaffected [211]. This may change, however, as feature sizes decrease further, and technological limits are approached.

The function of an interconnect or wiring system is to distribute clock and other signals and to provide power/ground, to and among, the various circuits/systems functions on a chip [96]. The ever-increasing integration densities of chips are accompanied by a quest for increasing speeds of signal transmissions over wires. Unfortunately, the expected delay of a wire of a certain length, which is proportional to the wire’s resistance as well as to its capacitance, tends to compare unfavorably to gate delays when feature sizes decrease. In practice, the increased delays of wires relative to gate delays under scaling are mostly felt by wires whose lengths cannot be downscaled, due to them running across the chip. Wires that do scale down in length are less sensitive (though not immune) to this problem [86]. Since most architectures require a substantial level of non-local interconnects, many wires will need to be sufficiently wide to offer the decreased delays associated with lower resistance. This has resulted in designs with an interconnection hierarchy on chip layouts of local thin wires at the bottom, and intermediate wires connecting functional units in the middle, to fat wires for global routing at the top of the hierarchy [192]. There is much incentive to develop new conductor and dielectric materials, to keep pace with technology requirements. Notable has been the replacement of aluminum by copper for conductors from around the late 1990’s, as well as the introduction of low- $\kappa$  dielectrics, which allow decreased pitch (interspacings) between wires.

Whereas transistor delay used to be the bottleneck in past technology generations, it has been superseded by wire delays. The reason for this lies in the increased needs for relatively long wires due to the increased complexities of designs [86]. The growing gap between transistor delays and wire delays can be addressed through the use in wires of *repeaters* at constant-length intervals to each other. Though this tends to result in smaller delays, it goes at the expense of increased silicon area and power consumption [86]. Ultimately the solution lies in shorter wires, and this favors architectures with local interconnection schemes [11]. Absent the use of such architectures, other radical concepts and solutions to the interconnect problem need to be contemplated. The ITRS of 2005 [96] mentions the following options:

1. Different signaling methods, like *Raised Cosine Signaling* [10], which uses raised cosine pulses instead of square pulses to cope with noise crosstalk problems. Also being investigated is *Resonant Clocking* [153], which uses on-chip inductors to effectuate resonance of clock pulses, and which leads to substantially decreased power dissipation.

2. Innovative design and package options that have optimization of the interconnects as their prime objectives.
3. Three-dimensional interconnects, to decrease the average wire length [166]. It may be a challenge to remove heat from such structures, however; it may also be necessary to use new system architectures and design tools [188].
4. Different physical principles by which to transfer signals. Among these, optical interconnects [139] promise high propagation speeds and bandwidths, high precision clock distribution, absence of crosstalk, and reduced power dissipation. Another possibility is signal transmission by Radio Frequency (RF) microwaves [151], which is accomplished through the placement of small antennas on a chip that are able to receive a global signal, like a clock signal. Finally, guided terahertz waves [104] and plasmons (e. g. [205]), which are hybrids of RF and optical signaling, use transmission frequencies around 1 THz and may result in significantly increased bandwidth.
5. Nanoelectronics-based solutions. Carbon nanotubes [106] are especially interesting due to their high conductance, and their potential to be used as semiconductors as well. Other solutions may include the use of conducting channels, such as molecular interconnects, and quantum-mechanical interactions, such as spin coupling and tunneling (see [203] for an overview).

Connecting wires to each other on nanometer-scales is nontrivial due to alignment issues. For this reason, wires are often interconnected through a perpendicular arrangement, like in crossbar arrays, which are discussed in more detail in Sect. “Crossbar Array-Based Architectures”.

The function of devices is the processing of signals input to them and, in response, outputting signals according to certain specifications. The Field-Effect Transistor (FET) has long been the basis for successive generations of CMOS. The generation commercially available from the end of 2007 – processor chips based on a 45 nm process – will feature high- $\kappa$  dielectrics and metal gates [22]. This combination allows for a thicker gate insulation layer given a certain gate capacitance, thus substantially reducing leakage through tunneling – the reason for this being that tunneling decreases exponentially with an increasing thickness of the insulation.

Key factors in the success of MOSFETs and circuits based on them have been noise tolerance of digital circuits and a good ability for fan-out due to the high signal gains of transistors [167], and these factors may also feature prominently in nano-electronics. Moreover, in order to be

competitive with CMOS, nano-electronic devices and circuits should satisfy the following requirements [95,189]:

- Scalable by several orders of magnitude beyond CMOS.
- High information/signal processing capacity. This may be achieved through high switching speeds or a high degree of parallelism.
- Energy dissipation that is much less than in CMOS.
- Room temperature operation.

Nanoelectronic devices also need to satisfy requirements specific to their small feature sizes. Alignment, for example will be much more difficult for three-terminal devices, like transistors, than if only two terminals are involved, like in FETs based on crossed nanowires [40].

Another important requirement for nanometer-scale devices is that they constitute a *universal set of operators*, which means that every desired circuit can be constructed as a circuit based on these devices. Most designers take it for granted that any arbitrary logic circuit can be constructed from p-type and n-type transistors, but this requirement may become less trivial when alternative devices are used. What then are promising candidates to become successors of MOSFETs?

Carbon Nanotube (CNT) FETs are considered to be the foremost candidate in this context. The channel through which current is rectified by a gate in CNT FETs consists of a single-walled carbon nanotube with semiconducting transport properties (e. g. see [4]). There are still many open issues with CNTs, including the need for a better understanding of the physical mechanisms underlying their operation, problems with the synthesis of CNTs with appropriate characteristics, and fabrication issues like placement, material integration, On/Off ratios, etc. [95].

Resonant tunneling devices [64,129] include resonant tunneling diodes (RTD), which have two terminals, and resonant tunneling transistors, with three terminals. An RTD consists of a double barrier structure through which electron transmission is highly efficient about certain resonance energy levels. These devices have novel characteristics, such as negative differential resistance, and they have potentially very high switching speeds and could result in circuits with a reduced number of components and power dissipation. However, they seem to be unable to form a universal set of operators, since no designs are known for memory cells and logic gates only involving RTDs [148].

Single Electron Tunneling (SET) devices [198] are based on the controlled motion of individual electrons. These devices involve junctions, through which tunneling takes place, and Coulomb islands that temporarily store electrons and that are usually implemented as quantum

dots. SET devices have the potential for high density and power efficiency, while allowing fast switching, but they tend to suffer from low noise immunity as well as from the problem that their gain is insufficient to allow satisfactory fan-out [95].

Molecular devices [134,159] utilize single molecules – typically small organic ones – acting as electronic switches and storage elements. Extremely high densities may be achieved with these devices, and they are thought to be very suitable for fabrication by chemical synthesis. Though efficient power usage is expected, these devices may offer switching speeds that are not much higher than those possible by CMOS. Experimental devices have been built, but it has not been made sufficiently clear yet whether the observed device properties are those of the molecules, or rather influenced by those of the used electrodes [95].

Spin logic devices [210], also referred to as magneto-electronics or spintronics, encode information by the magnetic spin orientations of electrons. Based on the manipulation of the magnetic spin of electrons by magnetic fields or by an applied voltage, these devices may allow very low power dissipation, but unfortunately their projected switching speeds may be limited.

Other devices that may play a future role in nanocomputers are Rapid Single Flux Quantum (RSFQ) devices [124], molecular-scale electromechanical devices [36,37], and quantum interference devices [46,213].

## Nanofabrication Techniques and Architectures

There is a close relationship between the technology used to manufacture computers and their architecture. The irregular, aperiodic, structures of current computers are only possible through the use of *top-down* microfabrication technology, like optical lithography (e.g. see [123,128]). Notwithstanding the physical limitations faced by lithography, its life-time has been extended over and over again, through for example resolution enhancement techniques [123], which can carry feature sizes below the illumination wavelength. Alternatives to optical lithography are nano-imprinting [34], in which a master with nanometer-scale features is made using techniques like electron-beam lithography. This master is then used as a stamp and pressed onto a (soft) target surface that can then be filled in with for example metal to create wires. Expected to allow manufacturing down to 10 nm, these techniques will impose relatively few restrictions on structures, though the number of times a master can be used is limited. As top-down fabrication techniques are being stretched to smaller feature sizes, they will increasingly impose restrictions on layouts, as recent de-

velopments involving resolution enhancement techniques show [123].

There is a lack of consensus as to what is beyond the year 2015 [94] in the context of top-down technologies, though – given the investments made in them – they are unlikely to disappear for the foreseeable future. That said, building the facilities for top-down fabrication will gradually run into economic limitations, as their costs tend to increase by a factor of two for every chip generation – a trend known as *Moore's second law*. Moreover, feature sizes of less than 10 nm may be intractable for top-down mass fabrication, necessitating additional techniques based on different principles. Denoted as *bottom-up*, these techniques exploit interactions between atoms and between molecules to create useful structures. Bottom-up manufacturing is expected to come at the expense of decreased complexity in architectures, because of the lesser degree of control over the manufacturing process. Architectures possible through bottom-up techniques can be divided in three classes: regular, random, and – a combination of these two – quasi-regular.

*Regular* architectures are characterized by a repetitive pattern of simple features that can not be changed after fabrication. Typically, regular architectures require the use of some form of molecular *self-assembly*, which is defined as the autonomous organization of components into patterns or structures without human intervention [206]. Self-assembly is related to, but not synonymous with, self-organization, the latter being mainly concerned with pattern formation, and the former mainly involving pre-existing components (separate or distinct parts of a disordered structure) that are designed with a certain outcome of the process in mind [206]. Self-assembly as a nanofabrication method offers a number of advantages [156]:

- it is an inherently parallel process,
- it can generate structures with sub-nanometer precision, and
- it can generate 3-dimensional structures.

When self-assembly's outcome is controlled through external forces or geometrical constraints that are supplementary to the original interactions driving it, it is called *directed self-assembly*. The control allowed by directed self-assembly extends from merely tuning the interactions between individual assembling components to positioning the components at a desired location [156], like the positioning of wires through an electrical field [177]. Manufacturing by self-assembly requires sufficient control over the process to guarantee regularity. This implies the existence of some mechanism that – though in itself lacking the control of top-down fabrication – can check and

correct errors in the fabrication process. The manufacturing of structures based on the self-assembly of DNA tiles [116,160,169,209] is an example of this. Such tiles form a skeleton that can be filled in later, using techniques to transport matter towards certain locations in DNA tiles [175]. Though practical architectures based on such tiles are yet to be realized, computing on such structures in the future can be envisioned. An important question in this context is how complex each tile should be. If its complexity approaches that of a conventional processor, it is unlikely that it can be manufactured by self-assembly, because of the absence of much regularity, even in the case the processor is not very complex. Self-assembly would require a much simpler tile, with a complexity not exceeding that of a simple finite automaton. This has led various researchers to investigate *Cellular Automata*, which are based on regular (usually 2-dimensional) arrays of tiles (cells), each filled with a finite automaton (► [Cellular Automata as Models of Parallel Computation](#)). Section “[Cellular Automaton-Based Architectures](#)” describes the use of this model for nanocomputer architectures.

Another regular architecture attracting attention from researchers is the *Crossbar array*. It employs a 2-dimensional mesh of wires that have switches at their crosspoints (see Sect. “[Crossbar Array-Based Architectures](#)”). Regular architectures require post-fabrication configuration to impose a certain desired functionality on them. Configuration may take place, for example, by software initializing cell states of a cellular automaton or setting the switches of a crossbar array. To allow reuse of the architecture for computation, one usually assumes that it can be configured more than once in various configurations. In other words, the architecture should be *reconfigurable*.

At the other end of the spectrum are *random* or *unstructured* architectures. The irregularity of such architectures tends to be mainly caused by the inability of the manufacturing process to perfectly control the formation of structures. The resulting randomness is usually considered part of the architecture to begin with. The elements that are not part of an unstructured architecture will be considered defects, because they are not planned. Lacking a structure imposed on them at manufacturing time, such architectures rely on post-fabrication configuration. To configure a circuit – which may for example be as complex as a processor – on the underlying hardware, it is necessary to have detailed information in advance about the hardware’s structure. This is usually accomplished by an external host computer. In other words, scanning of the underlying hardware is not self-contained in the architecture. This indicates that random structures will be mostly used in combination with one-time configuration of a cir-

cuit, since on-line scanning of the hardware by a host each time that a circuit needs to be mapped may be impractical and too time-consuming. Examples of random architectures are Neural Networks [197] and – related to them – the Nanocell [195] (see Sect. “[Neural Network-Based Architectures](#)”).

Randomness is hard to avoid in bottom-up fabrication of nanometer-scale features, due to the occurrence of defects. For this reason, perfectly regular architectures may be hard to achieve in nanocomputers. When defects can be dealt with by a one-time configuration of the hardware around them, the term *quasi-regularity* is used. See Sect. “[Fault-Tolerance](#)” for a discussion of defect-tolerance and fault-tolerance.

It is unlikely that one day mankind suddenly wakes up in a world dominated by bottom-up manufacturing. Rather, it will be gradually introduced, incorporated at first in the framework of top-down manufacturing, and increasing its share over the years. This will likely result in architectures in which CMOS is combined with nanoelectronics, like in [32,122,125,181,190,219].

## Heat Dissipation

The huge integration densities facilitated by nanotechnology will have far-reaching consequences for heat dissipation. Being a growing problem in VLSI chips, the heat dissipated by nanoelectronics per unit of area will reach impractical values if techniques, materials, devices, circuits, and architectures continue to be used as they have been in the past. Chips dissipate power proportional to the number of devices  $N$ , to the clock frequency  $f$ , to the probability  $p$  (typically 0.1) that a device is active during a clock cycle, and to the average energy dissipation  $E$  of an active device in one clock cycle [148,159], giving a power dissipation of  $P = NfpE$ . CMOS designs have been by and large governed by the scaling rules proposed in [51]. Table 2 shows some parameters in this context: the power dissipation per device decreases quadratically with downward scaling, and this is indeed what happened in the 30 years since the publication of the table. Coupled with a quadratically increasing number of devices per unit of area, this rule implies a constant power dissipation per unit area over successive generations. There is a limit in this regard, however, because the required downscaling of the power supply voltage causes an increased susceptibility to noise [102] and increased off-state drain leakage currents [137]. Ultimately the decreased switching energies associated with downscaling run into the thermal limit, discussed in Sect. “[Introduction](#)”. To make matters worse, the number of devices  $N$  and the frequency  $f$  will



### Nanocomputers, Table 2

Technological scaling rules for MOSFETs. From the 1970's on, each new chip generation – introduced at intervals of two to three years – has seen its minimum feature sizes reduced by a factor  $K \approx 0.7$ . The first item *device dimensions* applies to channel length, oxide thickness, gate width, etc. The table assumes unchanged electric fields in devices over successive generations

Device Parameter	Scaling Factor
Device dimensions	$1/K$
Voltage	$1/K$
Current	$1/K$
Gate capacitance	$1/K$
Delay time	$1/K$
Power dissipation	$1/K^2$

increase as feature sizes move into the range of nanometers. To see where this ultimately leads to, Zhirnov et al. take the physical parameters of an imaginary system to its extreme, resulting in a “limit technology”, in which a chip contains the smallest binary switches, packed to maximum density and operating at the lowest possible energy per bit. This would give a power density in the range of a few MW/cm<sup>2</sup> [218], which dwarfs the power density of the sun's surface by three orders of magnitude. Unfortunately, the capacity to remove heat is limited to several hundred W/cm<sup>2</sup> when known cooling methods for two-dimensional structures are used [218]. Additional measures thus seem justified, and this section looks for them in terms of architectures and algorithms.

Goldstein [74] proposes to reduce heat dissipation problems by spreading circuits out in space, reducing the number of times each part of a circuit is used in a computation. The resulting increase in parallelism allows for higher clock rates. Since the circuitry is spread over a larger area, a lower power density results, but this goes at the cost of longer wires, which causes increased delay and power consumption.

Other approaches seek to reduce heat dissipation in logic circuits by focusing on the clock. Each time clock signals are distributed over a circuit, energy will be pumped into wires, much of which gets lost in the form of heat. A novel approach to clocking problems is to use *resonant clocks*. These designs employ clock distribution circuitry that act as transmission lines with uniformly distributed capacitors and inductors, in which electromagnetic waves propagate long distances without the need for repeaters to amplify signals [9,31,141,152,212]. Since electrical charges on parasitic capacitors can be extracted and preserved by the inductors, this approach allows for a significantly reduced power consumption, but it has problems of its own, like unpredictability due to noise and reflections, diffi-

culties to distribute clock signals in a clock-tree due to impedance mismatches, and the inability to yield constant phase and magnitude of clock signals simultaneously [9]. To deal with some of these problems, while preserving the important advantage of low-power, Banu et al. [9] propose to use two identical side-by-side transmission lines with linear topologies in which clock signals travel in opposite directions, allowing the extraction of an absolute phase at any point of the clocking circuitry by averaging the two signals' timings. The overhead of the averaging circuitry, however, may compromise the efficiency of this solution.

A radical way to deal with clocking problems is to remove the clock all together, resulting in *asynchronous circuits*, which have a range of advantages. Apart from not having the clock's overhead of increased energy consumption and increased area, such circuits have the fortunate tendency to restrict energy consumption to only those parts that are actively switching. In other words, asynchronous circuits have an inherent mechanism to selectively shut off parts of them – even down to fine granularities of individual devices – a feature for which synchronous circuits require special *clock gating* circuitry.

Apart from the reduced power consumption and heat dissipation offered by asynchronous electronics – provided it is well-designed – there are also other advantages:

- Problems with the distribution of clock signals disappear, like *clock-skew* (timing differences at which different parts of a circuit receive the clock-signal) and *race conditions* (the failure of signals to reach their destinations within a clock cycle). These problems tend to get worse, otherwise, with increasing integration densities.
- Less noise and electromagnetic interference.
- Insensitivity to physical implementations and conditions. This implies more freedom in the timing of signals and layouts of circuits, which is useful in the reconfigurable architectures expected in nanocomputers [74].
- Average rather than worst-case performance. An asynchronous circuit operates as fast as switching times of devices allow, rather than being limited by the slowest parts of a circuit, which in synchronous systems imposes an upper limit to the global clock rate.
- Modularity. Asynchronous circuits can be designed in terms of modules that can be combined without considerations of timing restrictions. This facilitates flexible rearrangement of circuit elements into various circuits.

Asynchronous circuits come in many flavors [44,82], which are distinguished by the extent as to which assump-

tions are made about timing of signals, both of a circuit and in the interaction of a circuit with its environment. The most general class of asynchronous circuits assumes certain bounds on delays of signals [199], but unfortunately it is the least robust to unexpected behaviors, which complicates their design. Forming the second class are *self-timed circuits*, which only make assumptions on timing within circuit elements, like bounded delay etc., but not on timing between modules [174]. *Speed-independent circuits*, the third class, assume unbounded delays in circuit elements, but zero or negligible delays on wires between the elements [147]. The fourth and fifth class form the *quasi-delay-insensitive circuits* [130] and the *delay-insensitive circuits* [178]. The latter are robust to delays in both circuit elements as well as wires, but this is at the cost of significant design complexity. For this reason quasi-delay-insensitive circuits are more common. They differ from delay-insensitive circuits in their requiring a so-called *isochronic fork* [133], which is a fanout element of which all output branches have a similar delay. In practice, this class of circuits shares many similarities with speed-independent circuits.

Notwithstanding the advantages of asynchronous circuits, their use has been limited to few applications. This is partly due to the much smaller base of design and testing infrastructure when compared to synchronous technology, though asynchronous design support technology is increasingly being developed [191]. One of the key motivations to use synchronous circuits in the first place used to be ease of design, as there is – as compared to asynchronous circuits – less unexpected circuit behavior accompanying changes in the values of signals. The advantages of synchronous design have become less evident today than in the past, as timing problems have become more urgent with the increase of clock frequency, but synchrony still defines the mindset of generations of circuit designers.

The conditions surrounding nanoelectronics may change that. Not only will problems accompanying a clock become more severe with the increase of integration densities, the entire nature of systems will change. Whereas signals in digital logic have been voltage-encoded in virtually all systems, they may be represented in quite different ways on nanometer scales, to better reflect the physical environment. As signals more and more approach the scales of individual particles, a token-based representation in which signals have a discrete indivisible character may be more appropriate. Circuits based on single electron tunneling [198] are already moving in this direction [154], as do experimental circuits based on molecule cascades [58,85]. Asynchronous timing is very suitable for

token-based circuits, and may thus play an important role in this context [157].

Another approach to minimize heat dissipation aims to minimize the energy required for switching, even eliminate it altogether. This is possible – at least in theory – by making computations reversible [16,113] (► [Reversible Computing](#)). The key to this line of thought is the observation that the information stored on a computer is ultimately represented in terms of physical entities [115]. Operations in a computer that change the entropy of this information will accordingly change the thermodynamical entropy of the physical system. The erasure of a bit by overwriting it with a new value, is such an operation [113]. Whereas the bit's entropy corresponds to two possible (unknown) states of the bit before the operation, the bit's value will be fixed into one state by the operation, which decreases the entropy. This implies a decrease of the bit's physical degrees of freedom, and – to compensate for this – there is an increase of unobservable degrees of freedom such as in the microscopic motion of molecules [127], which usually translates into the dissipation of heat. The energy loss associated with such an operation would be at least  $kT \ln 2$  J, according to the thermal limit in Sect. “[Introduction](#)”. While Landauer [113] initially believed that no universal computations could be performed without a change in entropy, Bennett [16] proved this belief wrong by constructing a universal Turing-machine that conducts its operations reversibly, implying that it would cause no change in entropy.

This line of thought, which started in the early 1960's, is widely known today, but it has not resulted yet in practical circuits. Arguably, it will be difficult to overcome noise and friction, which are fundamental to physical processes. Most proponents of reversible computing agree that these factors are an impediment to dissipation-less computing, but they tend to argue that, absent these factors, reversible computing would *in principle* be possible. This would make it a matter of technological progress to approach the zero average switching energy limit arbitrarily close, rather than there being a hard physical lower limit.

One way to go about in this context is to slow down switching speed. Called *Adiabatic Switching* [5], this approach decreases the dissipation caused by the charging and discharging of a gate capacitance  $C$  through an associated resistance  $R$ . The dissipated energy for charging such a gate is given by [5]:

$$E = \frac{RC}{\tau} CV^2$$

whereby  $\tau$  is the time required to charge the gate, and  $V$  is the voltage supplied to the circuit. An increase in  $\tau$

thus implies a corresponding decrease in the dissipated energy, the other factors remaining equal. Schemes of adiabatic switching usually follow the requirement that, first, a switch is only closed when there is no voltage difference between its terminals and, second, a switch is only opened when no current flows through it. Adiabatic schemes of switching tend to be presented in combination with a reversible mode of computation [63]. When reversibility is left out, the resulting adiabatic scheme is usually characterized as *charge recovery* or *energy recycling* [173].

Reversible computing appears to have particular appeal for nanoelectronics, since physical interactions at nanometer scales are typically reversible. Notwithstanding the reversible computing paradigm's wide acceptance, there are still concerns about its feasibility [29,98,165]. Concerns based on problems with noise and friction have been pointed out in [165] in the early years of reversible computing. This is a recurrent theme with skeptics of reversible computation. The overwhelming responses generated by [165] are particularly educating [13,18,114,193], (see also the accompanying reporting in [168]). Other concerns [98] highlight some principal reasons why reversible computing would only work under unrealistic conditions, such as the necessity to have an infinite information-theoretic temperature of the scheme. There are also concerns [29] about possible problems when reversible schemes are combined with adiabatic schemes, the first supposedly being thermally isolated from the environment, while the second scheme is usually not. Other fundamental problems are pointed out in [29] as well with adiabatic switching: apart from thermal noise interfering with the reliability of a simple adiabatic bit-flipping operation, this noise is also responsible for a less than ideal charging curve of a circuit's capacitances, which causes an increased power consumption. Finally, a problem in [29] is claimed with the power supply of adiabatic circuits, which, in order to be the approximate ideal current source needed for adiabatic charging, would require a high internal resistance. The latter would cause much heat dissipation, not in the circuit itself, but in the power supply, so that the overall energy efficiency of the scheme would be inferior to non-adiabatic schemes. A more optimistic note about the prospects of reversible adiabatic computing is found in [63], though it is admitted that there remain significant challenges, the most fundamental of which are the requirement of switching devices that are less resistive and less leaky, and the requirement of higher quality power supplies and clocks, like the resonant clocks mentioned earlier in this section.

The problems with clocks in the framework of reversible computing have led some researchers to consider

whether reversible computing can be combined with asynchronous circuits, especially asynchronous circuits that are token-based and delay-insensitive [118,120,145]. Circuits, in which only one single signal token moves around at a time, do not need a clock, so they would qualify as reversible and asynchronous. Though computational universal circuits can be constructed in this way, they have obviously a problem in terms of efficiency. The alternative is circuits in which multiple tokens move around in parallel. This may lead, however, to situations in which reversibility is hard to define [118]. A circuit element having two or more tokens as asynchronous inputs for an operation will – in order to preserve a strict one-to-one relationship between input and output as defined through a (unitary) Hamiltonian operator [59] – need to monitor the arrival times of the individual tokens, but this requires a large overhead and it is counterintuitive to how a circuit is supposed to operate, suggesting that the underlying physical model may need reconsideration.

An unconventional strategy to reduce energy consumption is to use switching operations with a reduced reliability. Palem and co-workers [105,155] assert that by accepting a reduced reliability of switching, less energy will be required to conduct a switching operation. Palem [155] shows that if a switching operation is correct with a probability  $p(> 1/2)$ , then its energy consumption – assuming that the operation is irreversible – is bounded from below by  $kT \ln(2p)$ . A reduced reliability of switching limits the range of possible algorithms to those that accept some degree of randomness, like probabilistic algorithms [146], image processing algorithms, and classification algorithms (e.g. [66]). Somewhat related is a scheme proposed by Kish [103] that drives computations by noise. Shown is a simple model circuit containing a comparator with error rate around 0.5. Inspired by the highly efficient way in which information is processed in the brain, this thermal noise driven computing scheme requires a minimum energy requirement of  $1.1kT$  per bit. This scheme may require extensive error correction to make it work. The use of noise and fluctuations as a computational resource is also outlined in [42], as likely being a driving force for mechanisms in biological organisms [216]. Brownian motion provides a free search mechanism, allowing a molecule, signal, or other entity to explore its environment as it randomly moves around, thereby undergoing certain operations (like molecular binding) when it arrives at certain locations (like molecular binding sites). An early mentioning of Brownian motion for computations is made by Bennett in [17], who uses it for reversible computation in case thermal noise cannot be avoided. Bennett considers a Brownian implementation for a mechanical Turing

machine, in which signals move around randomly, searching their way through the machine's circuitry, only to be restricted by the circuit's topology. This search would, in principle, not require any energy to be provided from a power supply, the drawback being that the computation would move forward and backward evenly without any bias, and thus would take a long time to complete. Funneling the random motions into preferred directions would require the selective use of ratchets, but ratchets tend to consume energy, as shown by Feynman [61]. A mechanism somewhat similar to a Brownian Turing machine is found in nature, in the form of the Brownian tape-copying machine embodied by RNA polymerase [17]. The efficient – in terms of energy consumption – use of Brownian motion in various biological mechanisms, such as those involving Brownian molecular motors, suggest that biological organisms may provide inspiration for future nanocomputing systems in this respect.

Noise and fluctuations have also been used in the simulated annealing process of a Boltzmann machine that can be implemented by SET devices [215] (see also Sect. “Neural Network-Based Architectures”), though reductions in power consumption have not been the main purpose of this work.

## Fault-Tolerance

When a system experiences a failure in one of its parts and can continue to operate, possibly at a reduced level, rather than – as most systems would do – fail completely, it is called *fault-tolerant*. A fault-tolerant system may experience a reduced throughput or an increased response time in the face of one (or more) of its components failing. Fault-tolerance is an important property for nanocomputers, since the reliability of their components will likely be low due to:

- The probabilistic nature of interactions on nanometer scales, rooted in thermodynamics, quantum mechanics, etc.
- The individual behavior of particles, molecules, etc. gaining importance over their averaged behavior due to their decreased numbers, causing the “Law of Large Numbers” to become invalid below the device level [21,48].

When faults have a permanent character – either incurred at manufacturing time or later – they are usually referred to as *defects*. A nanocomputer architecture that can cope with defects is called *defect-tolerant*.

When faults have a transient character, i. e. when they occur during operation and are temporary, they are re-

ferred to as *transient faults*, or, simply, *faults*, like the more general term. They are mainly caused by environmental conditions, such as thermal noise, quantum fluctuations, and electromagnetic perturbations. Cosmic radiation, which is a major source of transient faults in CMOS technology, may be less of a concern, on the other hand, for nanoelectronics based on materials different from semiconductor silicon [72]. The term used to describe robustness of an architecture to the occurrence of transient faults is *fault-tolerance*, like the more general term.

Faults may also be *intermittent*, meaning that they occur during a certain (extended) time interval, then disappear, and may (or may not) later reappear. Intermittent faults tend to be caused by unstable or marginal hardware, due to for example manufacturing residues and process variations. These problems tend to increase with the increase in integration density [38]. Changes in operational parameters, such as temperature, voltage, and frequency changes may trigger intermittent faults [38]. Though transient faults and intermittent faults have many characteristics in common, the latter differ in that they tend to occur in bursts, at the same location. Repairing a part affected by intermittent faults usually eliminates them, whereas transient faults can not be prevented by repairs.

Tolerance to faults and defects can be accomplished at various levels of abstraction [76]: the *physical device level*, the *architectural level*, and the *application level*. An example relating to the physical device level is the tolerance of digital devices to noise due to the digital encoding of signals. Tolerance at the architectural level compensates for malfunctioning of devices by organizing them in structures with a certain level of *redundancy*. Such a structure may consist, for example, of replicated (identical) components, operating in parallel, on which majority voting is used to extract the result of an operation. Tolerance at the application level refers to the ability of a computing application to operate correctly despite defects and faults in the underlying hardware. Research aiming to improve the tolerance at the application level may be found in the framework of probabilistic algorithms or image processing algorithms that tolerate some noise in their input. As tolerance at the physical device level tends to decrease with increasing integration density, the tolerance at the architectural and application levels need to improve to compensate for this. It also works the other way around: it may make less sense, for example, to improve tolerance at the physical device level, if the application level is able to cope with a certain level of faults or defects on its own. Most research efforts on fault-tolerance in the nanocomputing field are in practice directed to the architectural level, and this section is no exception on that.



Redundancy in an architecture is achieved by equipping it with additional resources [112]. Especially important are the following types of redundancy:

- *Information redundancy* refers to the addition of information to the original data such as to be able to recover it in case of faults. *Error Correcting Codes (ECC)* are the usual form to represent redundant information (e. g. [207]). Used extensively to achieve reliable operation in communications and memory, error correcting codes provide a way to cope with the corruption of bits by encoding messages as code words that contain redundant information. This information is used to reconstruct the original code word in case errors occur.
- *Hardware redundancy* refers to the inclusion of additional hardware to the system. This may encompass the use of extra components working in parallel to mask any faults that occur – a technique called *static redundancy*, which includes the use of ECC. When only one element of the extra components are used at a time, to be switched out and replaced by a spare in case of a fault, the term *dynamic redundancy* is used. Dynamic redundancy is usually achieved through periodic testing of circuits or self-checking of circuits, followed up by the *reconfiguration* of a circuit, or even its *self-repair*.
- *Time redundancy* involves the repetition of operations in a system. This may take place after an error is detected, in which case a recovery procedure needs to be started up. Alternatively, as a standard procedure operations may be repeated a fixed number of times, after which the consecutive results are compared to each other. Time redundancy is only suitable to cope with non-permanent faults, since repetitions of operations by a defective component will only repeat incorrect results.

The different nature of defects as opposed to faults is reflected in the different strategies usually followed to make an architecture robust to them, though there tends to be an overlap.

The permanent nature of a defect implies that it needs to somehow be isolated from non-defective parts to prevent it from affecting a system's correct functioning. Strategies to cope with defects thus tend to first detect defects and then reconfigure the system around them. This requires spare components that become activated due to the reconfiguration, while the defective components (or defective connections) are made inaccessible. Absent reconfiguration, a system's performance will be adversely affected by a defect, in terms of throughput, speed, reliabil-

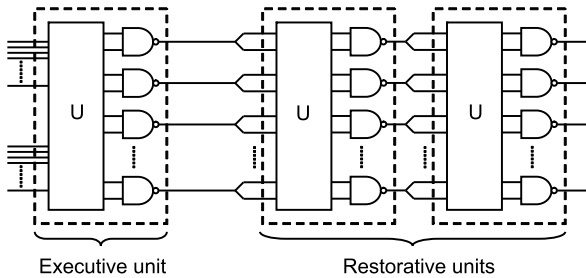
ity, etc., but this may be ameliorated to an acceptable level if sufficient spare resources are available.

Depending on whether reconfiguration takes place at the same time that a nanocomputer is processing data or not, the terms *on-line* or *off-line* reconfiguration, respectively, are used. On-line reconfiguration implies a continuous monitoring of a computer for defects and accordingly taking action if they are detected. Usually envisioned as being self-contained in the computer, this type of operation needs the detection of defects and the associated reconfiguration to be robust to defects and faults themselves, which complicates its realization. Off-line reconfiguration is usually done under the direction of a host-computer, which creates a defect-table and reconfigures the system accordingly. While easier to realize than on-line configuration, it may be only feasible if the size of a nanocomputer is limited, since otherwise the detection and reconfiguration operations would consume prohibitive amounts of time. A well-known example of off-line configuration of a prototype defect-tolerant nanocomputer is the Teramac [84]. Consisting of  $10^6$  gates operating at 1 MHz, it is based on Field Programmable Gate Arrays (FPGAs), 75 percent of which contain at least one defect. Having a total of three percent defects out of a total of 7,670,000 resources including wiring and gates, the Teramac needs a rich interconnection structure to find alternative routes around defects. To this end, it employs a crossbar array, which is described in more detail in Sect. "Crossbar Array-Based Architectures".

Fault-tolerance requires a different strategy than defect-tolerance, due to the temporary nature of faults. Such a strategy usually involves a combination of information redundancy and hardware redundancy.

Motivated by the low reliability of electronic components in his time, von Neumann studied how reliable computation could be obtained through the use of unreliable components [201]. His *multiplexing* technique employs a circuit built from faulty universal gates – all of the same type – to simulate a single gate of this type, but with increased reliability. Each (input- or output-) wire of the gate is replaced by a bundle of  $N$  wires in the circuit. The circuits investigated by von Neumann are based on the NAND-gate, but other gates could also be used in principle. The logic state 1 or 0 of a bundle of wires is determined by whether the fraction of wires in state 1 – called its *excitation level* – lies above resp. below a certain threshold. The scheme consists of three stages, of which the first – the *Executive Unit* – conducts the logic operation, and the last two – the *Restorative Units* – act as nonlinear amplifiers that restore the excitation level of the output bundle to its nominal level (see Fig. 1). The resulting circuit has





Nanocomputers, Figure 1

Von Neumann's fault-tolerant circuit that is used to replace a single faulty NAND-gate. It consists of three stages: The Executive Unit conducts the actual logic operation, whereas the two later Restorative Units reduce the fault-rate of the Executive Unit's output. The bundle of output wires of the circuit will effectively be reduced to one wire (not shown here), of which the value is determined by the excitation level of the bundle

two input bundles of wires and one output bundle of wires. Wires from the first input bundle are randomly paired by a permutation network labeled  $U$  with the wires from the second bundle; the resulting pairs of wires are then used as inputs to the faulty NAND-gates in the Executive Unit. The wires in the output bundle are split, and again randomly paired to form the inputs to the faulty NAND-gates in the first Restorative Unit. This is repeated for the second Restorative Unit. This scheme suffers from a limited reliability, while requiring a high redundancy. For example, an error rate of 0.005 per device requires a redundancy of 1000 to achieve a probability of failure of the gate of merely 0.027. A similar scheme by von Neumann that uses three-input *majority gates*, instead of NAND-gates, does not fare better.

A drawback of von Neumann's scheme is that it entirely focuses on the error rate of gates, but not on that of wires. While the latter can be modeled in terms of the former if wires have a constant length, this condition will not be valid in densely connected circuits, causing failure rates of gates to depend on the length of their input wires [163]. Von Neumann's scheme is at the root of many follow-up schemes proposed in the decades thereafter. Most of these efforts tend to minimize the number of required gates, given a certain failure probability of the components. Dobrushin and Ortyukov, for example, show in [54] that a function computed by  $m$  fault-free elements can be reliably computed by  $O(m \log m)$  unreliable elements with a high probability. This is further improved in [161,162] to  $O(m)$  unreliable elements. An excellent review of these models is found in [163]. This review also shows some variations on this line of research, in which it is not gates that may fail, but con-

tacts of switches. Thought to have become irrelevant with relays becoming obsolete for computational purposes, these variations regain significance in the context of nanocomputing, given that nanoscale switching devices with mechanical contacts have attracted renewed interest [36,37].

Another method, *R-fold modular redundancy* [52], achieves tolerance to faults by employing  $R$  copies of a unit ( $R$  preferably an odd number), of which the majority establishes the output in accordance with an operation conducted by a majority gate. If  $R = 3$ , this technique is called *Triple Modular Redundancy* (TMR) [52]. When the outputs of three TMR units are combined by a majority gate on a second level and so on in a hierarchy of levels, the result is a model with increased reliability higher in the hierarchy: *Cascaded Triple Modular Redundancy* (CTMR) [186]. TMR, CTMR, and von Neumann's multiplexing technique lead to high redundancies, as compared to reconfiguration techniques [149], but on the other hand they are more suitable for transient faults. Von Neumann's original multiplexing technique is reconsidered in [172], with claims that its redundancy can be reduced by using output wire bundles as input bundles to other gates, rather than – as von Neumann advocated – reducing them to a single wire between the outputs bundles of a gate-equivalent unit and the inputs bundles of the next unit. Von Neumann's multiplexing technique is applied in a hierarchical way in combination with reconfiguration techniques in [78]. The claim is of a significantly reduced redundancy in this design, even when the device error rate is up to 0.01. The usefulness of Markov chains in the analysis of these models is reviewed in [79].

The constructions of von Neumann and others have some resemblances with error correcting codes, especially to so-called *repetition codes*, which are codes (but not very efficient ones at that) in which each symbol of a message is repeated a number of times to create redundancy. The use of error correcting codes in this context leads to a scheme in which computation takes place in the encoded space, whereby errors are corrected locally, and encoding and decoding is only necessary at the beginning and the end, respectively, of the computation. This line of thought is followed in [187], but then through the use of *generalized Reed–Solomon codes* (e.g. [207]), rather than repetition codes, to realize a fault-tolerant computing model with improved reliability.

Fault-tolerance in architectures based on cellular automata and on crossbar arrays are discussed in the sections describing these architectures, i.e. in Sects. “[Cellular Automaton-Based Architectures](#)” and “[Crossbar Array-Based Architectures](#)”, respectively.

## Cellular Automaton-Based Architectures

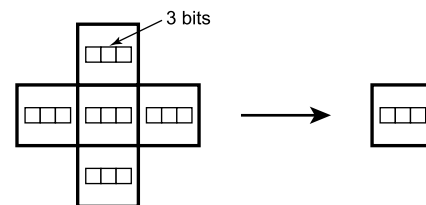
A Cellular Automaton (► [Cellular Automata as Models of Parallel Computation](#)) consist of cells that are organized in a regular array. Each cell contains a finite automaton that, taking as input the state of the cell and the states of its neighboring cells, changes its state according to a set of transition rules. These rules are designed such as to produce certain global behavior. When cellular automata are used as architectures for nanocomputers, this behavior is usually general-purpose computation, but application-specific behavior, like pattern recognition tasks, has also been investigated.

Originating in von Neumann's research on self-reproduction in the 1940's, cellular automata have experienced a resurgence in interest in recent years, especially in the context of nanocomputing architectures. Their simplicity – evidenced by a tiling-based structure in combination with local neighborhoods of cells with finite automata-like complexity – has much potential for bottom-up fabrication by molecular self-assembly. Moreover, their local interconnection structure implies a constant length of wires, which makes them efficiently scalable to high integration densities [11]. Cellular automata are unlikely to be used in combination with top-down fabrication methods, because of their significant overhead. Apart from the low percentage of cells that typically participate simultaneously in a computation, there may also be – depending on the design – relatively high hardware requirements for each cell, given its limited functionality as compared to conventional VLSI circuit designs. One of these factors may be improved, but this tends to be at the cost of efficiency with respect to the other factor: increasing the functionality of each cell, may increase the percentage of them that engage in a computation, since configurations of cells implementing certain tasks can be smaller, but on the other hand a cell will require more hardware resources to implement its functionality. A cellular automaton's overhead, which may exceed a factor of 10, could still be acceptable if it can be compensated for by the low-cost availability of cells in the huge quantities that can be expected with bottom-up fabrication. Even when only bottom-up fabrication methods are available, however, it is important to minimize the complexity of cells, because success in this area translates directly into more efficient physical realizations.

Though traditionally the complexity of a cellular automaton's cells is measured in terms of the number of cell states, this is inadequate for a nanocomputing framework, as it leaves unmentioned a cell's functionality in terms of the (number of) transition rules. The centralized storage

of the table of transition rules, which is common in traditional cellular automata or software to simulate them, is infeasible for nanocomputers, because it violates one of the key tenets of cellular automata, i. e., locality. That leaves only storage of the rule table in each cell individually as a viable option, implying that this table should be as small as possible, less a huge overhead occurs. The alternative to storage of the rule table is finding an efficient representation through inherently available physical interactions implied by the physical structure of a cell, and this is ultimately what many researchers aim for when using cellular automata for nanocomputers. This alternative, however, imposes even stricter limitations on transition rules.

A cell's complexity is determined by the information required to represent its functionality, and this includes the cell's state as well as the transition rules or any equivalent physical structure or mechanism. There are various ways to represent this information. A representation requiring a decreased number of bits for storage of transition rules may require increased resources, both in time and hardware, to decode the information. This implies that it is important to consider the overall costs, including the required supporting circuitry, to be taken into account when evaluating a cell's complexity. Absent a concrete physical implementation, it tends to be easiest to settle for a cost measurement in terms of the number of bits to represent the cell states plus the number of bits to represent the transition rules, under the assumption that decoding of this information is done through a straightforward and simple scheme. For the cellular automaton in [117], for example, three bits are required to represent a cell's five states. The von Neumann neighborhood of a cell in this model requires each transition rule to encode in its left-hand side the states of a cell itself and the states of four neighboring cells, and in the right-hand side the new state of the cell (see Fig. 2). Given that there are 58 transition rules in this model, this results in a total of



**Nanocomputers, Figure 2**

Bits required to encode a transition rule in a cellular automaton with 5-state cells and von Neumann neighborhood. The left-hand side of a transition rule requires three bits for the cell state itself and three bits each for the four neighboring cells' states. The right-hand side requires three bits to represent the new state of the cell after the transition

$3 + 58 \times 3 \times (1 + 4 + 1) = 957$  bits that are required to encode the functionality in a cell. A similar calculation yields 148 bits of information for the model in [157] and 100 bits of information for the model in [158]. A further reduction may be required to allow feasible physical realizations.

The use of cellular automata for nanocomputers originates in the late 1970's work of Fredkin, Toffoli, and Margolus, who aimed for computation inspired by physical models. A major issue in this research is the reduction of energy consumption as a road to high-performance high-integration computers. Signals in this paradigm have a discrete particle-like character and are conserved throughout operations. Moreover, a reversible scheme of logic (► [Reversible Cellular Automata](#)) is adopted.

One of the first proposals for a (2-dimensional) cellular automaton based on molecular interactions is due to Carter [28]. Written from a chemist's perspective, this proposal focuses on how particular operations of cells – including propagation in wires and switching by various automata – could be realized by means of the cascaded exchange of single and double bonds in molecules. This bond exchange – termed *soliton propagation in conjugated systems* – is relatively slow, but due to the small distances at which switches can be placed (in the order of 20 nm), switching times in the subnanosecond time scale are claimed to be feasible. Follow-up on this work is needed to show how the proposed operations could be combined into a cellular automaton capable of useful computations, but unfortunately the work has remained relatively unknown in the computer science community.

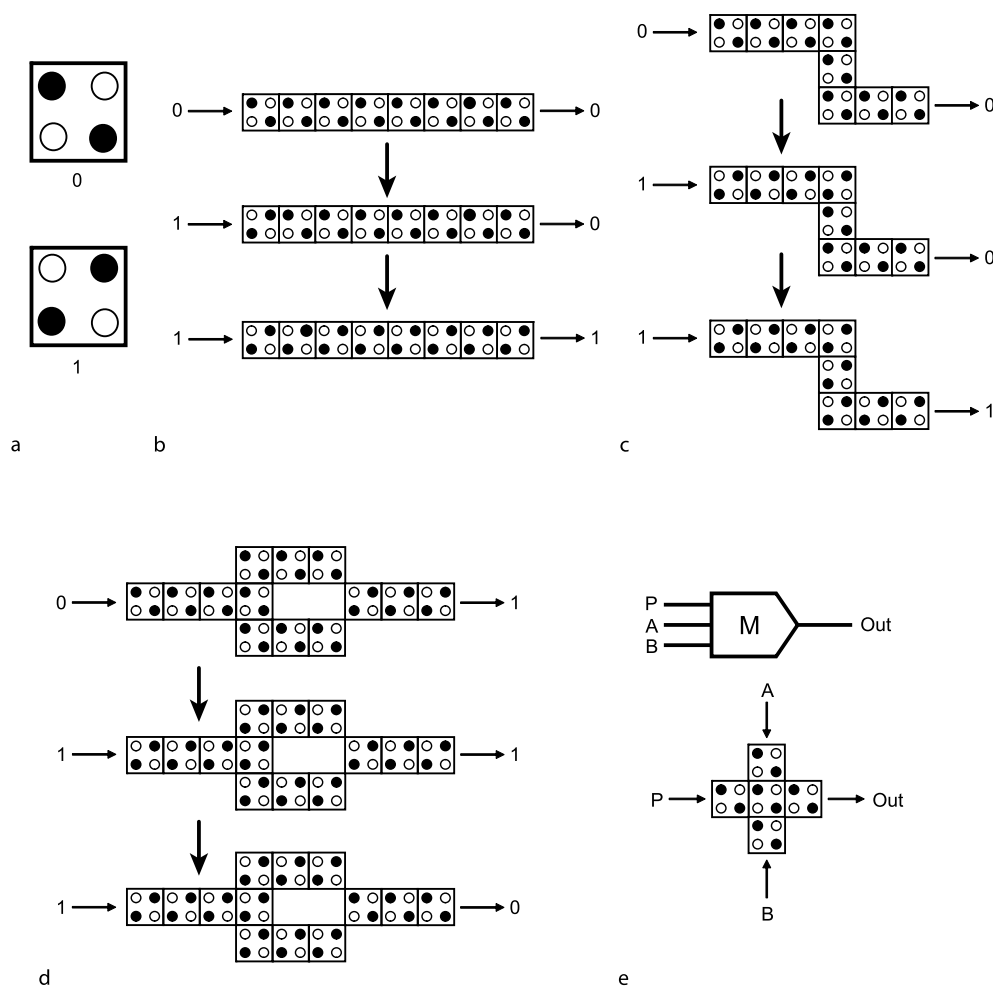
The Quantum Dot Cellular Automaton (QCA) [121, 164, 194] has cells that consist of four quantum dots, two of which contain an electron each (Fig. 3a). Due to Coulomb interactions inside a cell, the electrons will settle along one of two possible polarizations, interpreted as signals 0 and 1, respectively. On a larger scale, there are also Coulomb interactions between cells through the electrons in them. These interactions can be used for the propagation of signals along cells (Fig. 3b and c). Logic operations are also possible, like the NOT-gate in Fig. 3d, or the majority gate in Fig. 3e, which covers both an AND gate and an OR gate in its functionality. The QCA model is not a cellular automaton in the true sense, since it assumes a layout of cells along the topology of the circuitry they represent. This may even include layouts in which a cell is placed with an offset of half a cell at the side of another cell [194]. QCA promise extremely low power consumption as their operation does not involve flows of electrons, but rather tunneling of electrons on a local scale, i. e., between quantum dots within a cell. The problem with QCA, however,

is that they tend to be very sensitive to noise, imposing a low operating temperature on them, which is an obstacle to their efficient application. A supposedly more practical variation on QCA is the Magnetic QCA [39], in which the dots are implemented as submicron magnetic dots and which may allow operating temperatures as high as room temperature. There is considerable doubt among some researchers [24] whether QCA will ever provide a significant competitive advantage with respect to CMOS in terms of integration density, speed, power consumption, and the implementation of complex systems.

A 2-dimensional cellular automaton using optical signals to trigger transitions has been proposed by Biafore [20]. Employing physical interactions rather than explicit transition rules, this model has very simple partitioned cells that are structured such as to give control over both (a) which cells interact as neighbors and (b) when these interactions occur. Its design is based on Margolus' reversible cellular automaton [131] implementing Fredkin's billiard ball model (BBM) [65] (► [Reversible Cellular Automata](#)). The cells are suitable for implementation by quantum-dot devices that are organized such that electrical charge is transferred under the control of optical clock signals of different wave lengths supplied in a specific order. A similar way to control operations by electromagnetic or optical signals in a cellular automaton – though then applied to binary addition rather than universal computation – is described in [14]. Somewhat related is the cellular automaton model with less regular topologies in [15]. A 1-dimensional cellular automaton in which transitions are triggered by electromagnetic pulses of well-defined frequency and length is described in [126]. Apart from classical operations like the AND, OR, and NOT, quantum-mechanical operations are in principle possible by this scheme.

Cellular automata governed by nonlinear dynamics, so-called *Cellular Neural Networks* [35], hold another promise for nanocomputers, since the required physical implementations may be relatively simple. They are explored for this purpose in more detail in [101, 217]. The logic states of such cellular automata can be expressed in terms of the electrical phases in a dynamic physical process. A possible implementation of such a model is by *tunneling phase logic* [217]. Tunneling phase logic still suffers from many problems [23], however, ranging from the difficulty to manufacture uniform tunneling diodes to the occurrence of stray background charges.

Implementations of cellular automata utilizing physical interactions between CO molecules arranged on a copper (Cu) surface are described in [58, 85]. The atoms in the Cu surface form a triangular grid, on which the CO

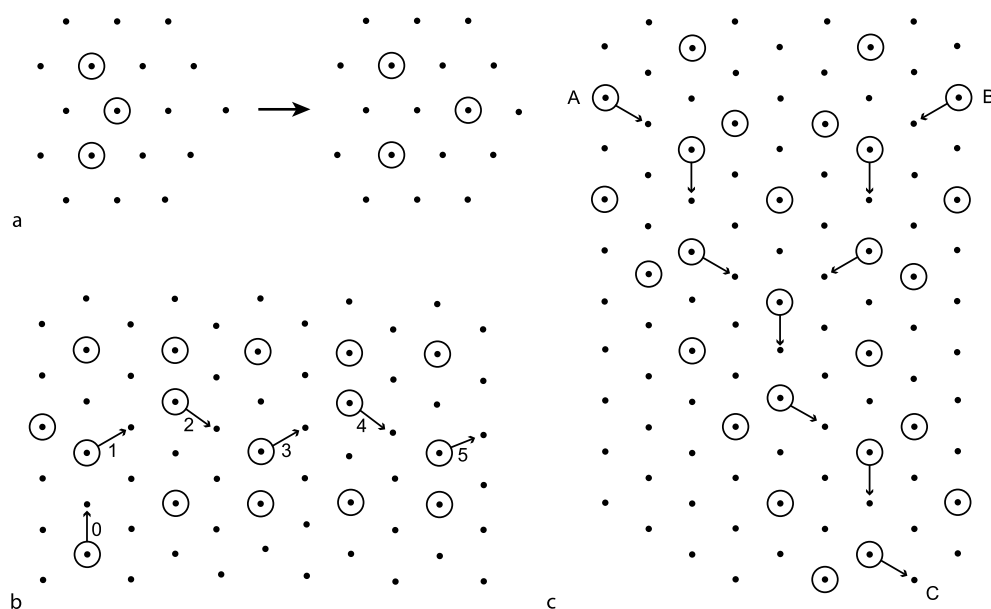


**Nanocomputers, Figure 3**

Quantum Dot Cellular Automaton (QCA). **a** A cell in the QCA contains four quantum dots storing two electrons. Since the electrons repel each other, they are likely to be in opposite corners of the cell. This allows for two possible cell configurations, which are interpreted as signals 0 and 1. **b** Due to the Coulomb interactions between electrons in quantum dots of neighboring cells, signals propagate along cells. A wire with cells in configuration 0 will change from the left to the right into a wire with cells in configuration 1 if the left-most cell is kept in configuration 1. **c** The same mechanism also works for wires taking left- and right-turns. **d** A NOT-gate formed from cells in the QCA. Change of input signal 0 to 1 results in the output signal changing from 1 to 0. **e** A majority-gate formed from cells in the QCA. From the majority-gate an AND-gate can be constructed by setting one of the inputs to the constant value 0. An OR-gate can be constructed by setting one input to 1

molecules settle, with the carbon atom of a molecule oriented toward a grid point. When two molecules are in nearest grid points, they will slightly point away from each other, due to a weak repulsive force between them. This force is sufficiently strong in some configurations of CO molecules to move certain CO molecules by one grid point, particularly in the so-called *Chevron-configuration* (Fig. 4a). Organized in suitable configurations, CO molecules will move on the grid in succession, triggering each other's hops, like domino stones falling in succession. These configurations, called *Molecule Cascades*, can

be exploited to transmit signals on the grid (Fig. 4b), and to conduct Boolean AND operations on signals (Fig. 4c). More complex circuits can also be formed, and these will require configurations to cross wires, and in some cases OR-gates. As compared to the interactions based on the Chevron configuration, the interactions required for wire crossings and OR-gates are more complicated, and unfortunately less reliable. Based on a wide variety of configurations, a three-bit sorter is even possible, as shown in [85]. A formal analysis of the configurations possible by molecule cascades is given in [25]. Though molecule cas-



Nanocomputers, Figure 4

Molecule Cascades based on CO molecules (indicated by circles) placed on a grid of copper atoms (indicated by dots). **a** Chevron configuration, in which the center molecule moves one grid point away due to repulsive forces. **b** Wire based on the Chevron configuration. The arrows indicate the hops made by the CO molecules. **c** AND-gate based on the Chevron configuration ( $C = A \wedge B$ )

acades allow impressive integration densities – the three-bit sorter would require a area of merely  $200 \text{ nm}^2$  – their speed of computation is very low: it takes about one hour for the three-bit sorter to complete its computation. There is no fundamental reason, though, why different systems of molecules would not be much faster. Setting up a molecular cascade system for computation involves a careful and time-consuming manipulation of individual molecules by a Scanning Tunneling Microscope (STM), so these types of systems are far from practical applications. A distinct disadvantage of the molecule cascades in [85] is that they allow only one-time computation: once a computation ends, the molecules have to be moved back to their initial positions to conduct the computation once more. An important step toward a practical system would be the development of a mechanism inherently present in the system to reinitialize molecules.

The cellular automata discussed above utilize the physical interactions inherent in their design to conduct transitions, an approach perhaps best illustrated in [132]. Though this tends to result in extremely simple cells, it may be challenging to build additional functionality into cells, such as functionality to configure cells into states necessary to initiate certain computations. This may be especially difficult if such a configuration process needs to take place on only part of the cell space. Likewise, error correction functionality is difficult to implement through

simple physical interactions in a cellular space. For these reasons some researchers advocate cells that are more complex, but still sufficiently simple to potentially allow a huge number of them to be fabricated by bottom-up techniques and organized in arrays at nanometer scale integration densities.

One approach along these lines is the *Cell Matrix* [57]. Each cell in this model, consisting of a memory of less than 100 bytes and a few dozen gates, can be configured to conduct an operation like NAND, XOR, one-bit addition, etc.

The above cellular automaton models are all timed synchronously, requiring the delivery of a central clock signal to each cell. A synchronous mode of timing causes a wide array of problems as pointed out in Sect. “Heat Dissipation”, like a high power consumption and heat dissipation. For this reason, cellular automata in which the principle of locality is carried one step further to the timing of the architecture, have started to attract interest in recent years. The resulting *asynchronous cellular automata* have a mode of operation in which all cells conduct their transitions at random times and independent of each other. Notwithstanding the advantage of asynchronous timing from a physical point of view, it brings up the question how to actually compute on such cellular automata in a deterministic way. The usual method to do this is to simulate a timing mechanism on an asynchronous cellular automaton to force the cells approximately into lock-step,

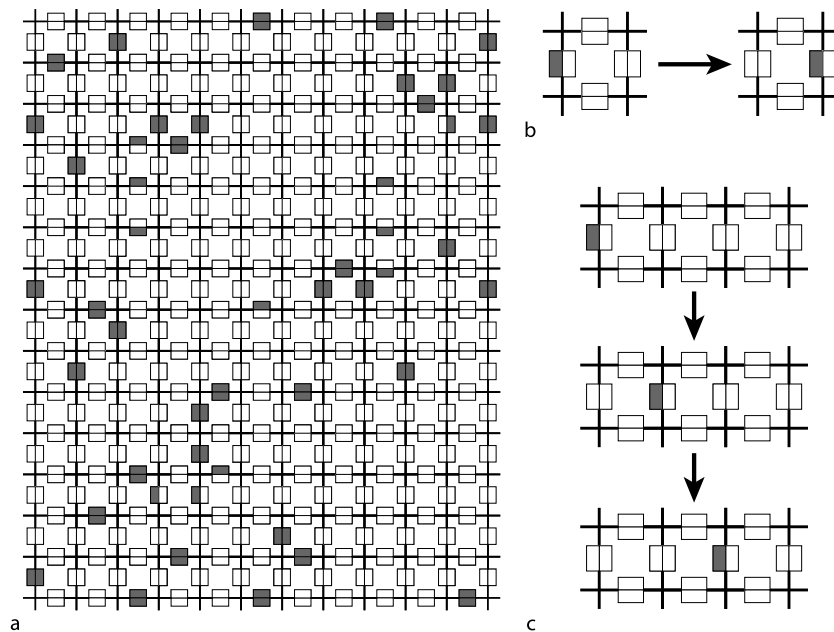


and then using well-established methods to compute synchronously. Unfortunately, such an approach causes a significant overhead in terms of the number of cell states and transition rules. To simulate an  $n$ -state synchronous cellular automaton on its asynchronous counterpart, one needs, depending on the method, a number of states of up to  $O(n^2)$  [157]. Worse yet, to synchronize different parts of an asynchronous cellular automaton with each other – necessary, because in the end it is a synchronous cellular automaton that is being simulated – exchange of signals between these parts is required. This exchange takes the form of so-called *synchronization waves* that propagate along the cell space [157]. Since all cells need to continuously change their states to let these waves pass, there are many transitions being conducted that do not contribute to the computation, even in areas of the cell space where no signals or active configurations are present. Implemented physically, such asynchronous cellular automata would thus need to consume much energy for dummy transitions, making them hardly better candidates for nanocomputer architectures than synchronous cellular automata.

A more efficient approach to conduct computations on asynchronous cellular automata is to conduct asynchronous computations directly, using only synchronization between cells on strictly local scales. This approach

is based on embedding a circuit on the cellular space and then simulating the circuit's operation through the cooperative transitions of the cells. While the circuits in synchronous models are typically standard logic circuits involving AND, OR, and NOT gates, different functionalities are required for asynchronous circuits, but the general idea of embedding circuits on the cell space remains the same. The approach in [1,117,119,157,158] is to use delay-insensitive circuits (see Sect. “Heat Dissipation”), which, due to their robustness to signal delays, combine very well with asynchronous cellular automata, since the requirement no longer holds that signals must arrive at certain locations at certain times dictated by a central clock, as in synchronous circuits. This takes away concerns about variations in the operational speed of cells and considerably simplifies the design of configurations representing circuit elements embedded in the cellular space.

To avoid the randomness in the timing of transitions interfering with proper deterministic operation of an asynchronous cellular automaton, its design involves transition rules that serialize transitions into well-controlled sequences of cell states. Some asynchronous cellular automata designed according to these principles can be found in [1,117,119]. Unfortunately, many of these models still require tens of transition rules, which may hinder



Nanocomputers, Figure 5

A Self-Timed Cellular Automaton (STCA) consists of cells with four partitions. All four partitions as well as one partition of each of the four neighboring cells are rewritten by transition rules. **a** Example of a small configuration of cells with partitions having values 0 (white) and 1 (black). **b** Transition rule for signal propagation. **c** Applying this rule to the configuration at the left results in a 1-partition moving to the right

efficient physical implementations. To this end, alternative models have been proposed, called *Self-Timed Cellular Automata (STCA)* [157], in which cells are subdivided in four partitions, each representing a part of a cell's state information (see Fig. 5a). This results in a substantial reduction of the number of transition rules. The model in [158] employs merely six rules. One of these rules, a rule for signal propagation is shown in Fig. 5b, and its effect on a cell configuration of a signal is shown in Fig. 5c.

In order to compute on a cellular automaton, it is necessary to configure it for the computation. One way to do this is by having a separate mechanism for (re-) configuration, through a separate layer of circuitry to address each individual cell like in a memory, and writing the required state into it. Another way is to have the reconfiguration done by the cells themselves. This requires the transition rules to contain sufficient functionality to represent the reconfiguration mechanism. Such functionality resembles self-reproduction of configurations in cellular automata. In other words, such cellular automata require *construction universality* (► [Self-Replication and Cellular Automata](#)). Unfortunately, it turns out that implementing construction universality on a cellular automaton tends to give a large overhead in terms of the number of transition rules. A self-contained mechanism for reconfiguration may thus not be practical, absent a novel ingenious way to implement construction universality, probably inspired by some biological mechanisms.

Reconfiguration in cellular automata can also be used to achieve defect-tolerance. In the approach in [90], defects are detected and isolated from non-defective cells through waves of cells in certain states propagating over the cell space, leaving defective cells, which are unable to adhere to the state changes, standing out to be detected as defective [90]. This approach is taken one step further in [93] by additionally scanning the cell-space for defect-free areas on which circuits can be configured. Being self-contained, this scanning process uses reconfiguration mechanisms resembling the self-reproduction functionality on cellular automata. An online approach to defect-tolerance is followed in [92], in which configurations called *random flies* float around randomly in the cellular space, and stick to configurations that are static. A configuration unable to compute due to defects is static in this model and will thus be isolated by a layer of random flies stuck to it. Being highly experimental, these approaches require an overhead in terms of the number of transition rules that may be too high to be of practical significance for nanocomputers.

Approaches to fault-tolerance in cellular automaton-based nanocomputers tend to focus on economic ways to

implement redundancy. Early work in this line of research is reported in [150]. This model can correct at most one error in 19 cells, but to this end each cell requires a neighborhood of 49 cells in its transition rules, which is much higher than usual cellular automata. The increased complexity of cells suggests that they will be very error-prone in physical implementations, limiting the practical value of this work. The model in [80] suffers from similar problems.

Better fault-tolerance is obtained in [67,68,70] with synchronous cellular automata, and in [69,204] with asynchronous cellular automata simulating synchronous cellular automata: the idea is to organize cells in blocks that perform a fault-tolerant simulation of a second cellular automaton, which on its turn is also organized in blocks, simulating even more reliably a third cellular automaton, and so on. This results in a hierarchical structure with high reliability at the higher levels, like with the CTMR technique mentioned in Sect. “[Fault-Tolerance](#)”. The cells in these models, however, are too complex to be suitable for efficient physical implementations, since they contain information regarding the hierarchical organization, such as block structure, address within a block, programs selecting transition rules, and timing information. In [91] a fault-tolerant STCA based on *BCH codes* (e.g. [207]) is proposed, but computation on this model is inefficient, since only one signal at a time is allowed.

Spielman's work [187] on implementing a computation scheme in an encoded space based on an error correcting code, mentioned in Sect. “[Fault-Tolerance](#)”, has resulted in related work [158] in the context of STCAs. Each partition of an STCA's cell and each adjacent partition of a neighboring cell is stored in a memory and encoded by an error correcting code. Up to one third of the memory's bits can – if corrupted – be corrected by this method.

Cellular automata in which cells have the complexities of processing elements [62,202], rather than (much simpler) finite automata, form the other end of the spectrum of cellular nanocomputer architectures. Since the fabrication of even simple processing elements are likely to be beyond the reach of bottom-up techniques, this approach may be infeasible for the same reasons as to why alternatives to von Neumann architectures are proposed in the first place for nanocomputers.

In conclusion, the main challenges for cellular automaton-based nanocomputers are:

- the design of models requiring a very small number of transition rules i. e., in the order of less than 10,
- the design or discovery of a mechanism to implement transition rules in a physically feasible way,

- to implement reconfiguration ability, while keeping the number of transition rules small,
- to implement fault- and defect-tolerance while keeping the number of transition rules and cell states small.

### Crossbar Array-Based Architectures

A crossbar consists of a set of, say,  $N$  horizontal wires that cross another set of  $M$  vertical wires. The  $N \times M$  cross points of the wires form a matrix of switches, which can be individually set (programmed) in a low-resistance state (closed) or a high-resistance state (opened). The simple structure of two planes of parallel wires allows for a wide variety of bottom-up fabrication techniques, such as self-assembly by fluidics-based alignment [87], directed self-assembly based on chemical techniques in combination with an alternating current (AC) electric field [53], or fabrication including top-down techniques like nano-imprinting [99,214]. Alignment of wires in order to connect them – a major problem in nanoelectronics – is relatively straightforward for crossbar arrays, due to the wires being perpendicular. Combined with a significant flexibility in making and altering connections, this accounts for much of the attention for crossbar arrays in the context of nanocomputers. Nanowires with diameters of only a few nanometers that can be used in implementations of crossbar arrays have already been fabricated, like single-crystal nanowires [41,87,100,144] and carbon nanotube wires [184]. The crossbar in [171] employs carbon nanotubes as its wires that can be programmed by applying a voltage across a crosspoint of two wires. This voltage, which is higher than that for normal (non-programming) operation of the crossbar, results in the wires being bent toward each other to close the switch. After removing the voltage, these deformations are preserved due to van der Waals forces, so memory functionality is inherent to this mechanism. To open the switch again, an opposite voltage is applied, which drives the wires apart. Nanometer scale implementations of crossbar arrays like the above provide significant efficiency as compared to their CMOS-based counterparts, the latter requiring a separate 1-bit memory for each cross point to store the state of the corresponding switch.

Another nanoscale crossbar design places a very thin layer of rotaxane molecules between the two wire planes in the crossbar [32]. Molecules in this layer at crosspoints can be set into low-resistance or high-resistance states – again through applying a voltage of sufficient strength – and these states tend to remain unchanged throughout successive read operations.

The crossbar is used in the following contexts:

**Routing** When used as an interconnection array, the crossbar array allows the routing of any input from a (single) vertical wire to any output on a (single) horizontal wire, by closing the switch between these wires. The resulting flexibility in connectivity provides an efficient way to route around defects.

**Logic** Logic when implemented by crossbar arrays usually takes the form of wired (N)OR-planes or (N)AND-planes [47,109,182,183], whereby wires in one plane provide inputs and wires in the other plane provide outputs.

**Memory** The 1-bit memories at the crosspoints offer the potential of huge storage capacities when combined in large crossbar arrays [32,77,107,214]. Reading out a memory bit is usually done by applying a voltage to one of the corresponding wires, resulting in the state of the memory appearing on the other wire. Realizations as associative memories have also been considered, whereby information storage is distributed over the switches such that recovery is possible even if some switches fail [45,179].

**Address Decoding** The connections between the two wire planes can also be set such that a demultiplexer is obtained, which can address a single wire out of the  $N$  wires in one plane by much less than  $N$  wires in the other plane, e. g. down to  $O(\sqrt{N})$  or even  $O(\log N)$  wires [32,47,208,219]. Apart from use in combination with crossbar arrays for routing, logic, or memory, a demultiplexer is also an excellent interface between micro-scales on one hand and nano-scales on the other, allowing for a limited number of microwires in one plane of a crossbar array to address a much larger number of much thinner nanowires in the other plane. Such demultiplexers tend to be static, i. e., they do not require reprogramming of the switch settings. When used as micro/nano-interfaces, reprogramming may even be impossible as nanowires can not be accessed directly; prepatterned connections [47] or stochastically assembled connections [208] may be suitable options in this case.

The cross points in crossbar arrays can take various forms, the most common of which are resistors, diodes – both combined with a switch in each cross point – and Field Effect Transistors (FETs). Resistor crossbars, followed by diode crossbars, are easiest to fabricate by bottom-up techniques, since both are 2-terminal devices, which significantly simplifies alignment problems. Resistor crossbars, additionally, can be built using metal nanowires, facilitat-

ing low output impedances, but their linear device characteristics may complicate the approximation of nonlinear functionalities [180]. Though diode and FET crossbars will do better with regard to the latter point, they carry significant disadvantages with existing fabrication technology [180]: limited current density, forward-biased voltage drops, relatively high impedance of semiconductor nanowires, difficulty of configuration, etc. Resistor and diode crossbars, on the other hand, come with another disadvantage [189]: being 2-terminal devices, they are passive, thus lacking signal gain – a major disadvantage as compared to 3-terminal devices like transistors – complicating the restoration of signals. Moreover, 2-terminal devices tend to have higher power consumption than devices like transistors, due to the inherently higher currents flowing to ground [189].

The lack of signal restoration is especially problematic when many levels of logic are required through which signals pass. Minimizing this number of levels, however, causes an inefficient use of hardware: it prevents the exploitation of the logic structure of a circuit design, since each level – such as an OR-plane or an AND-plane in a crossbar array – has a homogeneous structure. This is a well-known problem in Programmable Logic Arrays (PLAs) [49]: an  $n$ -input XOR, for example, requires a number of product terms that is exponential in  $n$  when expressed by the 2-level logic of a single PLA. It is thus unsurprising that signal restoration in crossbar arrays has received significant attention.

Several proposals have been made to include signal restoration functionality in crossbar arrays. DeHon [50] realizes some of the cross points between two wires as FETs, such that the electric field of one wire controls the conductivity of the other wire. This requires doping of (parts of) a wire, but the problems related to that, such as alignment of doped parts of wires with crossing wires, may require further research. Kuekes et al. [110] propose to use a bistable-switch latch to restore signals, in which the logic value of a wire is controlled by two switches, one connecting to a logic 0 and one to a logic 1, such that only one switch is closed at a time. Goldstein [75] also uses a latch for signal restoration, though in this case the latch is composed of a wire with two Negative Differential Resistance molecules at either end. There are also schemes in which the inherent poor voltage margins of resistor-based crossbar logic is improved by including error-correcting codes in the scheme [109,111].

Within the class of crossbar arrays for memory applications, there is a pronounced difference in terms of storage capacity between crossbar arrays based on diodes and FETs on the one hand, and crossbar arrays based on re-

sistors on the other hand. The use of diodes or FETs ensures that each cross point is independent of the other cross points, since electrical current can only flow in one direction in these designs. In resistor-based crossbar arrays, on the other hand, there will often be an indirect (parasitic) path through which a current flows when wires are connected at multiple cross points to perpendicular wires. This has a profound impact on the storage capacity of memories based on crossbar arrays. Whereas an  $N \times N$  crossbar array will have an  $O(N^2)$  storage capacity when based on diodes or FETs, it will have only  $O(N \log N)$  storage capacity when based on resistors (for more detailed estimates of storage capacities see [185]).

Though a standalone crossbar array is insufficient as an architecture for nanocomputers, it can be applied at different places and at different levels in a hierarchy to form a more or less “complete” architecture.

Goldstein’s NanoFabrics [75], for example, consist of so-called nanoBlocks and switch-blocks organized in tiles called clusters that are connected to each other through long nanowires of varying lengths (e.g. 1, 2, 4, and 8 clusters long). These wires allow signals to traverse a few clusters without the need to go through switches. The nanoBlocks themselves consist of logic based on crossbar arrays, and the switch-blocks consist of interconnections based on crossbar arrays. Defect-tolerance is obtained by testing the circuits – like in the Teramac, mentioned in Sect. “[Fault-Tolerance](#)” – but unlike the Teramac, testing is conducted in incremental stages in which an increasing part of the hardware – once it has been configured around defects – is used in the testing of the remaining parts [140]. Testing the initial part is conducted by a host.

DeHon’s Nanowire-Based Programmable Architecture [49] provides designs for Wired-OR logic, interconnection, and signal restoration implemented by atomic-scale nanowire crossbar arrays.

The CMOL architecture [125,190] is based on a reconfigurable crossbar array of nanodevices that is superimposed at a small angle on a grid of CMOS cells. It is used to implement memory, FPGA logic, or neural network inspired models (see Sect. “[Neural Network-Based Architectures](#)”). Each CMOS cell consists of an inverter and two pass transistors serving two pins that connect to the crossbar array. Snider et al.’s Field-Programmable Nanowire Interconnect (FPNI) [181] improves on this proposal by adjusting the crossbar array’s structure so that all CMOS-fabricated pins can be of the same height. Logic functionality in the architecture is restricted to the CMOS part and routing to the crossbar of nanowires.

Reconfigurability is one of the strong points of crossbar arrays, but it requires a time-intensive process of test-

ing for defects to guarantee defect-free routes. In effect, manufacturing yield is traded for an expensive die per die multi-step functional test and programming process [24]. A combination with error-correcting codes, like in [108], is thought to be more economical by some authors [7]. It is unclear as to what extent reconfiguration around defects can be made self-contained in crossbar-array-based architectures.

### Neural Network-Based Architectures

A *Neural Network* is a collection of threshold elements, called neurons, of which the human brain contains about  $10^{11}$ , each connected to  $10^3$  to  $10^4$  other neurons. Neurons receive inputs from each other through weighted connections. The strengths (weights) of the connections in a neural network are changed in accordance with the input to it; this process is called *learning*. Learning is usually accomplished through a *Hebbian learning rule*, which was first discovered by Hebb to be of fundamental importance in real brains. Hebbian learning can be shortly described as “neurons that fire together, wire together”. In other words, two neurons tend to strengthen the connection between them when their patterns of fired pulses correlate positively with each other. In artificial neural networks this rule is often implemented in terms of the correlation of the time-averaged analog values of neurons, rather than the actual pulses. Hebbian learning lies at the basis of an impressive self-organizing ability of neural networks. The model has much flexibility with regard to the connectivity of the network between the neurons, as a result of which little detailed information is required about the underlying network structure to conduct learning processes. This tends to result in a good tolerance to defects and faults, which is an important reason for the interest in neural networks for nanocomputer architectures.

Neural networks tend to be especially useful to problems that involve pattern recognition, classification, associative memory, and control. Recognition of visual images, for example, is conducted by brains to an extent that is unsurpassed by traditional computers, in terms of speed (taking approximately 100 ms), as well as in terms of fidelity. Though individual neurons are slow – it takes about 10 ms for a neural signal to propagate from one neuron to the other – the system as a whole is very fast. Contrast that to the time it takes computers with clock speeds of many GHz to performing the same task, which tends to be in the order of minutes to hours.

Most studied applications of artificial neural networks have difficulty to go beyond the toy-problem level. A better knowledge about the brain may lead to better results,

and there is also the expectation [125] that the availability of huge amounts of hardware resources may be of help. It is important to realize, however, that the large-scale organization of the brain is directed not only by relatively easy-to-characterize learning algorithms, but also by genetic factors, which involve a certain degree of “wetness”. The latter is not very compatible with the usual perceptions about computers. It is questionable whether the abundance of hardware resources alone will result in a breakthrough in neural networks-based architectures.

*Neuromorphic Engineering* is the term used to describe the interdisciplinary field that takes inspiration from neuroscience, computer science and engineering to design artificial neural systems. Though initially associated with VLSI systems engineering – due to Mead [135] coining the term *neuromorphic* – it is increasingly attracting contributions related to nanocomputers. One of the early papers in this context is [170], in which quantum dots arranged in a two-dimensional array and deposited on a resonant tunneling diode (RTD) are connected by nanowires, to form nodal points of nonlinear bistable elements in a conductive network. The dynamics of this system bears strong similarities to the equations representing associative memories.

Another proposal related to neuromorphic nanoarchitectures is the *Nanocell*. A Nanocell is an element containing randomly ordered metallic islands that are interlinked with molecules – acting as switches – between micrometer-sized metallic input/output leads [196]. To impose certain functionalities on the nanocell, like an AND, a NOR, an Adder, etc., it is necessary for these switches to be set into certain configurations. Genetic Algorithms (GA) are used for this in [195], but this method requires a detailed knowledge of the random structure inside the nanocell, so it is impractical, resulting in proposals to use neural networks, in which – ideally – the switches can be programmed through a learning process employing electrical pulses on input and output terminals of the nanocell [88]. To create useful circuits from nanocells it is necessary to connect them – after programming – into certain circuits. Creating such inter-cell connectivity has not been researched in detail, but neuromorphic learning methods may be suitable in this context as well.

Boltzmann machine neural networks based on single electron tunneling devices are proposed in [215]. *Simulated annealing*, the processing in Boltzmann machines through which a minimum energy state is found in an optimization problem, is implemented through a digital oscillator based on the stochastic nature of the tunneling phenomenon. The architecture consists of a network in which outputs of neural elements are fed back to their in-



puts through weights in a interconnection network, which are implemented by capacitance values reflecting the particular problem to be optimized.

*CrossNets* [197] are neuromorphic architectures based on the CMOL architecture. Neurons in this model, being relatively sparse, are represented by CMOS logic, and the neural interconnections, being very dense, are represented by a crossbar array of nanometer scale wires, whereby a cross-connection between two wires represents a so-called *synapse*. The crossbar array offers a very flexible framework to connect neurons to each other: it allows connectivity ranging from hierarchical structures that are typically used in multi-layered perceptrons [83] to more homogeneous networks with quasi-local connections. CrossNets can be used for pattern classification in feedforward multilayered perceptrons as well as for pattern restoration in associative memories. The crossbar array, though very flexible, has also disadvantages in this context, because it limits interconnection weights to be binary, and it is hard to avoid interference between different switches when setting or updating weights. Some solutions to these problems may be reached by using multiple-wire cross connections per weight [197]. The above models employ neural signals that have analogous values, as opposed to the spiking signals that are common in real brains. Spikes offer the opportunity for decreased power consumption, and, more importantly, their timing appears to encode important information in brains, and is thought to allow the semantic binding of different brain areas with each other. Some preliminary results on spiking neuromorphic architectures based on CMOL CrossNets and comparisons to non-spiking models are given in [71]. A different neuromorphic architecture based on crossbar arrays is explored in [45] for use in associative memories.

## Future Directions

In pondering the future directions nanocomputers may take, it is important to realize that CMOS technology is a formidable competitor. Detailed predictions on the pace at which this technology progresses have been made for decades, with varying degrees of success. When the emphasis is on technological limitations, the resulting predictions tend to be overly conservative, as they do not take into account the many ways often found around those limitations [86]. When on the other hand trends are simply extrapolated from the past, predictions tend to be on the optimistic side, as they ignore hard physical limitations that might still be ahead [86]. Notwithstanding the uncertainties of the life-time of CMOS technology, there seems

to be a consensus that it can be extended to at least the year 2015. The exponential growth experienced over so many decades, however, may significantly slow down towards the end [143]. Beyond this, it is expected [23] that the benefits of increased integration densities will be extended for 30 more years through the use of emergent technologies. These technologies are likely to be used in combination with silicon technologies and gradually replace more of the latter.

Many of the limitations that are perceived in technology – and CMOS is no exception to this – are likely to be circumvented by reconsidering the underlying assumptions, as was observed by Feynman [60]. This extends to the framework of circuits and architectures as well. Reconsiderations of underlying assumptions has frequently resulted in paradigm shifts in computer science. A notable example form *Reduced Instruction Set Computers (RISC)* [73], which use a small set of rapidly executed instructions rather than a more expanded and slower set of instructions as was common until the 1980's in *Complex Instruction Set Computers (CISC)*. The CISC approach was advantageous in the context of limited sizes of memories, necessitating a dense storage of instructions at the cost of increased processing in the processor itself. That changed with the increasing sizes and decreasing costs of memories, coupled with the realization that only a small percentage of the instructions available in CISC processors were actually used in practice. In a similar way, a changing context will determine the direction in which architectures of nanocomputers develop. This chapter has identified a number of factors of importance in this respect, like the growing problems faced in the fabrication of arbitrary structures, the increasing occurrence of defects in those structures, the increasing occurrence of faults during computations, and the increasing problems with heat dissipation. These problems will have to be addressed one way or the other, and architectural solutions are likely to play an important role in them; regular or random structures in combination with (re)configuration techniques seem especially attractive in this context.

A gradual move away from the traditional von Neumann architecture toward the increasing use of parallelism is likely to occur. Initially it will be conventional processors that will be placed in parallel configurations – a trend that has already started in the first decade of the 21st century. As feature sizes decrease further, fabrication technology will be more and more bottom-up, resulting in parallelism becoming increasingly fine-grained, to the point that computers consist of a huge number of elements with extremely simple logic functionality that can be configured according to specifications by a programmer. In the end,

architectures may look more like intelligent memory than like the computers known today. Programming methods, though apt to change as architectures change, may be relatively unaffected by the trend toward finely grained parallelism, witness the resemblance of conventional programming languages to languages for fine-grained hardware platforms like FPGAs, some of which have been designed with resemblance to the programming language C in mind.

Trends in technology are often highly influenced by breakthroughs, and it is entirely possible that future nanocomputers develop in different directions. The increasing research efforts into unconventional computing offers a clue to the wide range of what is possible in this regard. Computation models based on new media like reaction-diffusion systems, solitons, and liquid crystals [2] are just a few of the fascinating examples that may some day work their way into practical computer architectures.

## Bibliography

- Adachi S, Peper F, Lee J (2004) Computation by asynchronously updating cellular automata. *J Stat Phys* 114(1/2): 261–289
- Adamatzky A (2002) New media for collision-based computing. In: *Collision-Based Computing*. Springer, London, pp 411–442
- Adleman LM (1994) Molecular computation of solutions to combinatorial problems. *Science* 266(11):1021–1024
- Appenzeller J, Joselevich E, Hönlein W (2003) Carbon nanotubes for data processing. In: *Nanoelectronics and Information Technology*. Wiley, Berlin, pp 473–499
- Athas WC, Svensson LJ, Koller JG, Tzartzanis N, Chou EYC (1994) Low-power digital systems based on adiabatic-switching principles. *IEEE Trans Very Large Scale Integr Syst* 2(4):398–407
- Aviram A, Ratner MA (1974) Molecular rectifiers. *Chem Phys Lett* 29(2):277–283
- Bahar RI, Hammerstrom D, Harlow J, Joyner WH Jr, Lau C, Marculescu D, Orailoglu A, Pedram M (2007) Architectures for silicon nanoelectronics and beyond. *Computer* 40(1):25–33
- Ball P (2006) Champing at the bits. *Nature* 440(7083):398–401
- Banu M, Prodanov V (2007) Ultimate VLSI clocking using passive serial distribution. In: *Future Trends in Microelectronics: Up the Nano Creek*. Wiley, Hoboken, pp 259–276
- Bashirullah R, Liu W (2002) Raised cosine approximation signalling technique for reduced simultaneous switching noise. *Electron Lett* 38(21):1256–1258
- Beckett P, Jennings A (2002) Towards nanocomputer architecture. In: Lai F, Morris J (eds) *Proc. 7th Asia-Pacific Computer Systems Architecture Conf. ACSAC'2002 (Conf. on Research and Practice in Information Technology)*, vol 6. Australian Computer Society, Darlinghurst, Australia
- Benioff P (1980) The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines. *J Stat Phys* 22(5): 563–591
- Benioff P (1984) Comment on: Dissipation in computation. *Phys Rev Lett* 53(12):1203
- Benjamin SC, Johnson NF (1997) A possible nanometer-scale computing device based on an adding cellular automaton. *Appl Phys Lett* 70(17):2321–2323
- Benjamin SC, Johnson NF (1999) Cellular structures for computation in the quantum regime. *Phys Rev A* 60(6):4334–4337
- Bennett CH (1973) Logical reversibility of computation. *IBM J Res Dev* 17(6):525–532
- Bennett CH (1982) The thermodynamics of computation – a review. *Int J Theor Phys* 21(12):905–940
- Bennett CH (1984) Thermodynamically reversible computation. *Phys Rev Lett* 53(12):1202
- Bennett CH (1988) Notes on the history of reversible computation. *IBM J Res Dev* 32(1):16–23
- Biafore M (1994) Cellular automata for nanometer-scale computation. *Physica D* 70:415–433
- Birge RR, Lawrence AF, Tallent JR (1991) Quantum effects, thermal statistics and reliability of nanoscale molecular and semiconductor devices. *Nanotechnology* 2(2):73–87
- Bohr MT, Chau RS, Ghani T, Mistry K (2007) The high  $\kappa$  solution. *IEEE Spectr* 44(10):23–29
- Bourianoff G (2003) The future of nanocomputing. *Computer* 36(8):44–53
- Brillouët M (2007) Physical Limits of Silicon CMOS: Real Showstopper or Wrong Problem? In: *Future Trends in Microelectronics; Up the Nano Creek*. Wiley, Hoboken, pp 179–191
- Carmona J, Cortadella J, Takada Y, Peper F (2006) From molecular interactions to gates: a systematic approach. In: *ICCAD '06: Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design*, San Jose, 5–9 Nov 2008
- Carter FL (1983) The chemistry in future molecular computers. In: *Computer Applications in Chemistry*, Proc. 6th Int. Conf. on Computers in Chemical Research and Education. Elsevier, Amsterdam, pp 225–262
- Carter FL (1983) Molecular level fabrication techniques and molecular electronic devices. *J Vac Sci Technol B* 1(4):959–968
- Carter FL (1984) The molecular device computer: point of departure for large scale cellular automata. *Physica D* 10(1–2):175–194
- Cavin RK, Zhirnov VV, Hutchby JA, Bourianoff GI (2005) Energy barriers, demons, and minimum energy operation of electronic devices. In: *Proc. SPIE*, vol 5844, pp 1–9
- Ceruzzi P (1998) *A history of modern computing*. MIT Press, Cambridge
- Chan SC, Shepard KL, Restle PJ (2005) Uniform-phase uniform-amplitude resonant-load global clock distributions. *IEEE J Solid-State Circuits* 40(1):102–109
- Chen Y, Jung GY, Ohlberg DAA, Li X, Steward DR, Jeppesen JO, Nielsen KA, Stoddard JF, Williams RS (2003) Nanoscale molecular-switch crossbar circuits. *Nanotechnology* 14(4):462–468
- Choi H, Mody C (2007) Molecular electronics in the *longue durée*: the microelectronics origins of nanotechnology. In: *Joint Wharton-Chemical Heritage Foundation Symposium on the Social Studies of Nanotechnology*, Philadelphia, 7–8 Jun 2007
- Chou SY, Krauss PR, Renstrom PJ (1996) Imprint lithography with 25-nanometer resolution. *Science* 272(5258):85–87
- Chua LO, Yang L (1988) Cellular neural networks: theory. *Circuit Syst IEEE Trans* 35(10):1257–1272

36. Collier CP, Wong EW, Belohradský M, Raymo FM, Stoddart JF, Kuekes PJ, Williams RS, Heath JR (1999) Electronically configurable molecular-based logic gates. *Science* 285(5426):391–394
37. Collier CP, Mattersteig G, Wong EW, Luo Y, Beverly K, Sampaio J, Raymo FM, Stoddart JF, Heath JR (2000) A [2]Catechane-based solid state electronically reconfigurable switch. *Science* 289(5482):1172–1175
38. Constantinescu C (2007) Impact of intermittent faults on nanocomputing devices. In: Workshop on Dependable and Secure Nanocomputing. Edinburgh, 28 Jun 2007
39. Cowburn RP, Welland ME (2000) Room temperature magnetic quantum cellular automata. *Science* 287(5457):1466–1468
40. Cui Y, Lieber CM (2001) Functional nanoscale electronic devices assembled using silicon nanowire building blocks. *Science* 291(5505):851–853
41. Cui Y, Lieber C, Lauhon L, Gudixsen M, Wang J (2001) Diameter-controlled synthesis of single crystal silicon nanowires. *Appl Phys Lett* 78(15):2214–2216
42. Dasmahapatra S, Werner J, Zauner KP (2006) Noise as a computational resource. *Int J Unconv Comput* 2(4):305–319
43. Davari B (1999) CMOS technology: present and future. In: *Proc. IEEE Symp. on VLSI circuits. Digest of Technical Papers*, pp 5–9
44. Davis A, Nowick SM (1997) An introduction to asynchronous circuit design. Tech Rep UUCS-97-013, Computer Science Department, University of Utah
45. Davis BA, Principe JC, Fortes JAB (2004) Design and performance analysis of a novel nanoscale associative memory. In: *Proceedings of 4th IEEE Conference on Nanotechnology*, pp 314–316
46. Debray P, Raichev OE, Rahman M, Akis R, Mitchel WC (1999) Ballistic transport of electrons in T-shaped quantum waveguides. *Appl Phys Lett* 74(5):768–770
47. DeHon A (2003) Array-based architecture for FET-based nanoscale electronics. *IEEE Trans Nanotechnol* 2(1):23–32
48. DeHon A (2004) Law of large numbers system design. In: *Nano, quantum and molecular computing: implications to high level design and validation*. Kluwer, Norwell, pp 213–241
49. DeHon A (2005) Nanowire-based programmable architectures. *ACM J Emerg Technol Comput Syst* 1(2):109–162
50. DeHon A, Lincoln P, Savage JE (2003) Stochastic assembly of sublithographic nanoscale interfaces. *IEEE Trans Nanotechnol* 2(3):165–174
51. Dennard RH, Gaensslen FH, Yu HN, Rideout VL, Bassous E, LeBlanc AR (1974) Design of ion-implanted mosfets with very small physical dimensions. *IEEE J Solid-State Circ* 9(5):256–268
52. Depledge PG (1981) Fault-tolerant computer systems. *IEE Proceedings A* 128(4):257–272
53. Diehl MR, Yaliraki SN, Beckman RA, Barahona M, Heath JR (2002) Self-assembled deterministic carbon nanotube wiring networks. *Angewandte Chem Int Ed* 41(2):353–356
54. Dobrushin RL, Ortyukov SI (1977) Upper bound for the redundancy of self-correcting arrangements of unreliable functional elements. *Probl Inform Transm* 13(3):203–218
55. Drexler KE (1986) *Engines of creation*. Anchor Books, New York
56. Drexler KE (1992) *Nanosystems: molecular machinery, manufacturing, and computation*. Wiley, New York
57. Durbeck LJK, Macias NJ (2001) The cell matrix: an architecture for nanocomputing. *Nanotechnology* 12(3):217–230
58. Eigler DM, Lutz CP, Crommie MF, Mahoran HC, Heinrich AJ, Gupta JA (2004) Information transport and computation in nanometer-scale structures. *Phil Trans R Soc Lond A* 362(1819):1135–1147
59. Feynman RP (1985) Quantum mechanical computers. *Optics News* 11:11–20
60. Feynman RP (1992) There's plenty of room at the bottom (reprint of 1959 lecture). *J Microelectromech Syst* 1(1):60–66
61. Feynman RP, Leighton R, Sands M (2006) Ratchet and pawl. In: *The Feynman Lectures on Physics*, vol 1. Addison Wesley, San Francisco, pp 1–9
62. Fountain TJ, Duff MJB, Crawley DG, Tomlinson CD, Moffat CD (1998) The use of nanoelectronic devices in highly parallel computing systems. *IEEE Trans VLSI Syst* 6(1):31–38
63. Frank MP (2005) Introduction to reversible computing: motivation, progress, and challenges. In: *CF '05: Proceedings of the 2nd conference on Computing frontiers*. ACM Press, New York, pp 385–390
64. Frazier G, Taddiken A, Seabaugh A, Randall J (1993) Nanoelectronic circuits using resonant tunneling transistors and diodes. In: *Digest of Technical Papers. IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, 24–26 Feb 1993, pp 174–175
65. Fredkin E, Toffoli T (1982) Conservative logic. *Int J Theor Phys* 21:219–253
66. Fukú H (2002) Nondeterministic density classification with diffusive probabilistic cellular automata. *Phys Rev E* 66(6):066106.
67. Gács P (1986) Reliable computation with cellular automata. *J Comput Syst Sci* 32(1):15–78
68. Gács P (1989) Self-correcting two-dimensional arrays. In: *Micali S (ed) Randomness in Computation. Advances in Computing Research (a scientific annual)*, vol 5. JAI Press, Greenwich, pp 223–326
69. Gács P (1997) Reliable cellular automata with self-organization. In: *IEEE Symposium on Foundations of Computer Science*, pp 90–99
70. Gács P, Reif X (1988) A simple three-dimensional real-time reliable cellular array. *J Comput Syst Sci* 36(2):125–147
71. Gao C, Hammerstrom D (2007) Cortical models onto CMOL and CMOS – architectures and performance/price. *IEEE Trans Circ Syst I: Regul Pap* 54(11):2502–2515
72. Gil D, de Andrés D, Ruiz JC, Gil P (2007) Identifying fault mechanisms and models of emerging nanoelectronic devices. In: *Workshop on Dependable and Secure Nanocomputing (DSN'07)*. Online proceedings [www.laas.fr/WDSN07/WDSN07\\_files/Texts/WDSN07-POST-01-Gil.pdf](http://www.laas.fr/WDSN07/WDSN07_files/Texts/WDSN07-POST-01-Gil.pdf). Accessed 5 Aug 2008
73. Gimarc CE, Milutinovic VM (1987) A survey of RISC processors and computers of the mid-1980s. *Computer* 20(9):59–69
74. Goldstein SC (2005) The impact of the nanoscale on computing systems. In: *IEEE/ACM International Conference on Computer-Aided Design (ICCAD 2005)*. San Jose, CA, pp 655–661. Online Proceedings [www.cs.cmu.edu/~seth/papers/goldstein-iccad05.pdf](http://www.cs.cmu.edu/~seth/papers/goldstein-iccad05.pdf). Accessed 5 Aug 2008
75. Goldstein SC, Budiu M (2001) Nanofabrics: Spatial computing using molecular electronics. In: *Proceedings of the 28th annual international symposium on Computer architecture*, pp 178–191

76. Graham P, Gokhale M (2004) Nanocomputing in the presence of defects and faults: a survey. In: *Nano, Quantum and Molecular Computing*. Kluwer, Boston, pp 39–72
77. Green JE, Choi JW, Boukai A, Bunimovich Y, Johnston-Halperin E, Delonno E, Luo Y, Sheriff BA, Xu K, Shin YS, Tseng HR, Stoddart JF, Heath JR (2007) A 160-kilobit molecular electronic memory patterned at  $10^{11}$  bits per square centimeter. *Nature* 445(7126):414–417
78. Han J, Jonker P (2003) A defect- and fault-tolerant architecture for nanocomputers. *Nanotechnology* 14(2):224–230
79. Han J, Gao J, Qi Y, Jonker P, Fortes JAB (2005) Toward hardware-redundant, fault-tolerant logic for nanoelectronics. *IEEE Des & Test Comput* 22(4):328–339
80. Harao M, Noguchi S (1975) Fault tolerant cellular automata. *J Comput Syst Sci* 11(2):171–185
81. Hartmanis J (1995) On the weight of computations. *Bull Eur Assoc Theor Comput Sci* 55:136–178
82. Hauck S (1995) Asynchronous design methodologies: an overview. *Proc IEEE* 83(1):69–93
83. Haykin S (1998) *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ
84. Heath JR, Kuekes PJ, Snider GS, Williams RS (1998) A defect-tolerant computer architecture: Opportunities for nanotechnology. *Science* 280(5370):1716–1721
85. Heinrich AJ, Lutz CP, Gupta JA, Eigler DM (2002) Molecule cascades. *Science* 298(5597):1381–1387
86. Ho R, Mai KW, Horowitz MA (2001) The future of wires. *Proc IEEE* 89:490–504
87. Huang Y, Duan X, Wei Q, Lieber C (2001) Directed assembly of one-dimensional nanostructures into functional networks. *Science* 291(5504):630–633
88. Husband CP, Husband SM, Daniels JS, Tour JM (2003) Logic and memory with nanocell circuits. *IEEE Trans Electron Dev* 50(9):1865–1875
89. Hush NS (2003) An overview of the first half-century of molecular electronics. *Ann N Y Acad Sci* 1006:1–20
90. Isokawa T, Abo F, Peper F, Kamiura N, Matsui N (2003) Defect-tolerant computing based on an asynchronous cellular automaton. In: *Proceedings of SICE Annual Conference, Fukui, Japan*, pp 1746–1749
91. Isokawa T, Abo F, Peper F, Adachi S, Lee J, Matsui N, Mashiko S (2004) Fault-tolerant nanocomputers based on asynchronous cellular automata. *Int J Mod Phys C* 15(6):893–915
92. Isokawa T, Kowada S, Peper F, Kamiura N, Matsui N (2006) On-line marking of defective cells by random flies. In: Yacoubi SE, Chopard B, Bandini S (eds) *Lecture Notes in Computer Science*, vol 4173. Springer, Berlin, pp 347–356
93. Isokawa T, Kowada S, Takada Y, Peper F, Kamiura N, Matsui N (2007) Defect-tolerance in cellular nanocomputers. *New Gener Comput* 25(2):171–199
94. International Roadmap Committee (2005) *International Technology Roadmap for Semiconductors*
95. International Roadmap Committee (2005) *International Technology Roadmap for Semiconductors, Emerging Research Devices*. [www.itrs.net/Links/2005ITRS/ERD2005.pdf](http://www.itrs.net/Links/2005ITRS/ERD2005.pdf). Accessed 5 Aug 2008
96. International Roadmap Committee (2005) *International Technology Roadmap for Semiconductors, Interconnect*. [www.itrs.net/Links/2005ITRS/ERD2005.pdf](http://www.itrs.net/Links/2005ITRS/ERD2005.pdf). Accessed 5 Aug 2008
97. Iwai H (2004) CMOS scaling for sub-90 nm to sub-10 nm. In: *VLSI '04: Proceedings of the 17th International Conference on VLSI Design*, IEEE Computer Society, Washington, DC, p 30
98. Jablonski DG (1990) A heat engine model of a reversible computation. *Proc IEEE* 78(5):817–825
99. Jung GY, Johnston-Halperin E, Wu W, Yu Z, Wang SY, Tong WM, Li Z, Green JE, Sheriff BA, Boukai A, Bunimovich Y, Heath JR, Williams RS (2006) Circuit fabrication at 17nm half-pitch by nanoimprint lithography. *Nano Lett* 6(3):351–354
100. Kamins TI, Williams RS, Chen Y, Chang YL, Chang YA (2000) Chemical vapor deposition of Si nanowires nucleated by TiSi<sub>2</sub> islands on Si. *Appl Phys Lett* 76(5):562–564
101. Kiehl RA (2006) Information processing in nanoscale arrays: DNA assembly, molecular devices, nano-array architectures. In: *ICCAD '06: Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design*, San Jose, 5–9 Nov 2006
102. Kish LB (2002) End of Moore's law: thermal (noise) death of integration in micro and nano electronics. *Phys Lett A* 305(3–4):144–149
103. Kish LB (2006) Thermal noise driven computing. *Appl Phys Lett* 89(14):144104
104. Knap W, Deng Y, Rumyantsev S, Lu JQ, Shur MS, Saylor CA, Brunel LC (2002) Resonant detection of subterahertz radiation by plasma waves in a submicron field-effect transistor. *Appl Phys Lett* 80(18):3433–3435
105. Korkmaz P, Akgul BES, Palem KV, Chakrapani LN (2006) Advocating noise as an agent for ultra-low energy computing: probabilistic complementary metal-oxide-semiconductor devices and their characteristics. *Jpn J Appl Phys* 45(4B):3307–3316
106. Kreup F, Graham AP, Liebau M, Duesberg GS, Seidel R, Unger E (2004) Carbon nanotubes for interconnect applications. In: *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, pp 683–686
107. Kuekes PJ, Williams RS, Heath JR (2000) Demultiplexer for a molecular wire crossbar network. *US Patent* 6 128 214
108. Kuekes PJ, Robinett W, Seroussi G, Williams RS (2005) Defect-tolerant interconnect to nanoelectronic circuits: internally redundant demultiplexers based on error-correcting codes. *Nanotechnology* 16(6):869–881
109. Kuekes PJ, Robinett W, Williams RS (2005) Improved voltage margins using linear error-correcting codes in resistor-logic demultiplexers for nanoelectronics. *Nanotechnology* 16(9):1419–1432
110. Kuekes PJ, Steward DR, Williams RS (2005) The crossbar latch: Logic value storage, restoration, and inversion in crossbar circuits. *J Appl Phys* 97(3):034301
111. Kuekes PJ, Robinett W, Roth RM, Seroussi G, Snider GS, Williams RS (2006) Resistor-logic demultiplexers for nanoelectronics based on constant-weight codes. *Nanotechnology* 17(4):1052–1061
112. Lala PK (2001) *Self-checking and fault-tolerant digital design*. Morgan Kaufmann, San Francisco, CA
113. Landauer R (1961) Irreversibility and heat generation in the computing process. *IBM J Res Dev* 5(3):183–191
114. Landauer R (1984) Dissipation in computation. *Phys Rev Lett* 53(12):1205
115. Landauer R (1992) Information is physical. In: *PhysComp'92: Workshop on Physics and Computation*, Dallas, 2–4 Oct 1992, pp 1–4



116. Le J, Pinto Y, Seeman NC, Musier-Forsyth K, Taton TA, Kiehl RA (2004) DNA-templated self-assembly of metallic nanocomponent arrays on a surface. *Nano Lett* 4(12):2343–2347
117. Lee J, Adachi S, Peper F, Morita K (2003) Embedding universal delay-insensitive circuits in asynchronous cellular spaces. *Fundamenta Informaticae* 58(3/4):295–320
118. Lee J, Peper F, Adachi S, Mashiko S (2004) On reversible computation in asynchronous systems. In: *Quantum Information and Complexity*. World Scientific, Singapore, pp 296–320
119. Lee J, Adachi S, Peper F, Mashiko S (2005) Delay-insensitive computation in asynchronous cellular automata. *J Comput Syst Sci* 70:201–220
120. Lee J, Peper F, Adachi S (2006) Reversible logic elements operating in asynchronous mode. US Patent 6 987 402
121. Lent CS, Tougaw PD, Porod W, Bernstein GH (1993) Quantum cellular automata. *Nanotechnology* 4(1):49–57
122. Li C, Fan W, Lei B, Zhang D, Han S, Tang T, Liu X, Liu Z, Asano S, Meyyappan M, Han J, Zhou C (2004) Multilevel memory based on molecular devices. *Appl Phys Lett* 84(11):1949–1951
123. Liebmann LW (2003) Layout impact of resolution enhancement techniques: impediment or opportunity? In: *Proc. 2003 Int. Symp. on Physical Design (ISPD'03)*, ACM Press, pp 110–117
124. Likharev KK, Semenov VK (1991) RSFQ logic/memory family: a new Josephson-junction technology for sub-terahertz-clock-frequency digital systems. *IEEE Trans Appl Supercond* 1(1):3–28
125. Likharev KK, Strukov DB (2005) Introduction to Molecular Electronics. In: Cuniberti G et al (ed) *CMOL: Devices, circuits, and architectures*. Springer, Berlin, pp 447–477
126. Lloyd S (1993) A potentially realizable quantum computer. *Science* 261(5128):1569–1571
127. Lloyd S (2000) Ultimate physical limits to computation. *Nature* 406(6799):1047–1054
128. Madou MJ (2002) Lithography. In: *Fundamentals of Microfabrication, The Science of Miniaturization*. CRC Press, Florida, pp 1–76
129. Maezawa K, Förster A (2003) Quantum transport devices based on resonant tunneling. In: *Nanoelectronics and Information Technology*, pp 407–424
130. Manohar R, Martin AJ (1995) Quasi-delay-insensitive circuits are turing-complete. Tech. Rep. CaltechCSTR:1995.cs-tr-95-11, California Institute of Technology, Pasadena, CA
131. Margolus NH (1984) Physics-like models of computation. *Physica D* 10(1/2):81–95
132. Margolus NH (1999) Crystalline computation. In: *Feynman and computation: exploring the limits of computers*. Perseus books, Cambridge, pp 267–305
133. Martin AJ (1990) Programming in VLSI: From communicating processes to delay-insensitive circuits. In: Hoare CAR (ed) *Developments in Concurrency and Communication*. Addison-Wesley, Reading, pp 1–64
134. Mayor M, Weber HB, Waser R (2003) Molecular Electronics. In: *Nanoelectronics and Information Technology*. Wiley, Berlin, pp 501–525
135. Mead C (1990) Neuromorphic electronic systems. *Proc IEEE* 78(10):1629–1636
136. Mead C, Conway L (1980) Introduction to VLSI Systems. Addison-Wesley, Boston
137. Meindl JD (1995) Low power microelectronics: retrospect and prospect. *Proc IEEE* 83(4):619–635
138. Meindl JD, Chen Q, Davis JA (2001) Limits on silicon nanoelectronics for terascale integration. *Science* 293(5537):2044–2049
139. Miller DAB (2000) Rationale and challenges for optical interconnects to electronic chips. *Proc IEEE* 88(6):728–749
140. Mishra M, Goldstein SC (2003) Defect tolerance at the end of the roadmap. In: *Proceedings of the IEEE International Test Conference (ITC)*, vol 1, pp 1201–1210
141. Mizuno M, Anjo K, Surni Y, Wakabayashi H, Mogami T, Horiuchi T, Yamashina M (2000) On-chip multi-ghz clocking with transmission lines. In: *2000 IEEE International Solid-State Circuits Conference (ISSCC)*. Digest of Technical Papers, pp 366–367
142. Montemerlo MS, Love JC, Opitck GJ, Goldhaber-Gordon DJ, Ellenbogen JC (1996) Technologies and designs for electronic nanocomputers. Tech. Rep. 96W0000044, MITRE
143. Moore GE (2003) No exponential is forever: but “forever” can be delayed! In: *Solid-State Circuits Conference*. Digest of Technical Papers. ISSCC. IEEE International Solid-State Circuits Conference (ISSCC), vol 1, pp 20–23
144. Morales A, Lieber C (2001) A laser ablation method for the synthesis of crystalline semiconductor nanowires. *Science* 291(5348):208–211
145. Morita K (2003) A simple universal logic element and cellular automata for reversible computing. *Lect Notes Comput Sci* 2055:102–113
146. Motwani R, Raghavan P (1995) *Randomized Algorithms*. Cambridge University Press, New York, NY
147. Muller DE, Bartky WS (1959) A theory of asynchronous circuits. In: *Proceedings of an International Symposium on the Theory of Switching*. Harvard University Press, pp 204–243
148. Nikolić K, Forshaw M (2003) The current status of nanoelectronic devices. *Int J Nanosci* 2(1/2):7–29
149. Nikolić K, Sadek A, Forshaw M (2002) Fault-tolerant techniques for nanocomputers. *Nanotechnology* 13(3):357–362
150. Nishio H, Kobuchi Y (1975) Fault tolerant cellular spaces. *J Comput Syst Sci* 11(2):150–170
151. O KK, Kim K, Floyd B, Mehta J, Yoon H, Hung CM, Bravo D, Dickson T, Guo X, Li R, Trichy N, Caserta J, Bomstad W, Branch J, Yang DJ, Bohorquez J, L Gao L, Sugavanam A, Lin JJ, Chen J, Martin F, Brewer J (2003) Wireless communications using integrated antennas. In: *Proc. 2003 IEEE International Interconnect Technology Conference*, San Francisco, 2–4 June 2003, pp 111–113
152. O'Mahony F, Yue CP, Horowitz MA, Wong SS (2003) A 10-GHz global clock distribution using coupled standing-wave oscillators. *IEEE J Solid-State Circ* 38(11):1813–1820
153. O'Mahony F, Yue CP, Horowitz M, Wong SS (2003) 10 GHz clock distribution using coupled standing-wave oscillators. In: *Solid-State Circuits Conference*. Digest of Technical Papers. IEEE International Solid-State Circuits Conference (ISSCC), vol 1, pp 428–504
154. Ono Y, Fujiwara A, Nishiguchi K, Inokawa H, Takahashi Y (2005) Manipulation and detection of single electrons for future information processing. *J Appl Phys* 97:031101
155. Palem KV (2005) Energy aware computing through probabilistic switching: a study of limits. *IEEE Trans Comput* 54(9):1123–1137



156. Parviz BA, Ryan D, Whitesides GM (2003) Using self-assembly for the fabrication of nano-scale electronic and photonic devices. *IEEE Trans Adv Packag* 26(3):233–241
157. Peper F, Lee J, Adachi S, Mashiko S (2003) Laying out circuits on asynchronous cellular arrays: a step towards feasible nanocomputers? *Nanotechnology* 14(4):469–485
158. Peper F, Lee J, Abo F, Isokawa T, Adachi S, Matsui N, Mashiko S (2004) Fault-tolerance in nanocomputers: a cellular array approach. *IEEE Trans Nanotechnol* 3(1):187–201
159. Petty M (2007) *Molecular Electronics, from Principles to Practice*. Wiley, West Sussex
160. Pinto YY, Le JD, Seeman NC, Musier-Forsyth K, Taton TA, Kiehl RA (2005) Sequence-encoded self-assembly of multiple-nanocomponent arrays by 2D DNA scaffolding. *Nano Lett* 5(12):2399–2402
161. Pippenger N (1985) On networks of noisy gates. In: 26th Annual Symposium on Foundations of Computer Science, 21–23 October 1985, Portland, Oregon, IEEE, pp 30–38
162. Pippenger N (1989) Invariance of complexity measures for networks with unreliable gates. *J ACM* 36(3):531–539
163. Pippenger N (1990) Developments in: “The synthesis of reliable organisms from unreliable components”. In: *Proc. of Symposia in Pure Mathematics*, vol 50. pp 311–324
164. Porod W (1998) Quantum-dot cellular automata devices and architectures. *International journal of high-speed electronics and systems* 9(1):37–63
165. Porod W, Grondin RO, Ferry DK (1984) Dissipation in computation. *Phys Rev Lett* 52(3):232–235
166. Rahman A, Reif R (2000) System-level performance evaluation of three-dimensional integrated circuits. *IEEE Trans Very Large Scale Integr Syst* 8(6):671–678
167. Robert RW, Keyes W (1985) What makes a good computer device? *Science* 230(4722):138–144
168. Robinson AL (1984) Computing without dissipating energy. *Science* 223(4641):1164–1166
169. Rothemund PW, Papadakis N, Winfree E (2004) Algorithmic self-assembly of DNA sierpinski triangles. *PLoS Biol* 2(12):2041–2053
170. Roychowdhury VP, Janes DB, Bandyopadhyay S, Wang X (1996) Collective computational activity in self-assembled arrays of quantum dots: a novel neuromorphic architecture for nanoelectronics. *IEEE Trans Electron Dev* 43(10):1688–1699
171. Rueckes T, Kim K, Joselevich E, Tseng G, Cheung C, Lieber C (2000) Carbon nanotube based nonvolatile random access memory for molecular computing. *Science* 289(5476):94–97
172. Sadek AS, Nikolić K, Forshaw M (2004) Parallel information and computation with restitution for noise-tolerant nanoscale logic networks. *Nanotechnology* 15(1):192–210
173. Sathe V, Chueh JY, Kim J, Ziesler CH, Kim S, Papaefthymiou M (2005) Fast, efficient, recovering, and irreversible. In: *CF '05: Proceedings of the 2nd Conference on Computing Frontiers*. ACM, New York, pp 407–413
174. Seitz CL (1980) System timing. In: Mead CA, Conway LA (eds) *Introduction to VLSI Systems*. Addison-Wesley, Boston
175. Sherman WB, Seeman NC (2004) A precisely controlled DNA biped walking device. *Nano Lett* 4(7):1203–1207
176. Shor PW (2004) Progress in quantum algorithms. *Quantum Inf Process* 3(1–5):5–13
177. Smith PA, Nordquist CD, Jackson TN, Mayer TS, Martin BR, Mbindyo J, Mallouk TE (2000) Electric-field assisted assembly and alignment of metallic nanowires. *Appl Phys Lett* 77(9):1399–1401
178. Snepscheut JvD (1985) Trace theory and VLSI design. In: *Lecture Notes in Computer Science*, vol 200. Springer, Berlin
179. Snider GS, Kuekes PJ (2003) Molecular-junction-nanowire-crossbar-based associative array. US Patent 6 898 098
180. Snider GS, Robinett W (2005) Crossbar demultiplexers for nanoelectronics based on n-hot codes. *IEEE Trans Nanotechnol* 4(2):249–254
181. Snider GS, Williams RS (2007) Nano/CMOS architectures using a field-programmable nanowire interconnect. *Nanotechnology* 18(3):1–11
182. Snider GS, Kuekes PJ, Williams RS (2004) CMOS-like logic in defective, nanoscale crossbars. *Nanotechnology* 15(8):881–891
183. Snider GS, Kuekes PJ, Hogg T, Williams RS (2005) Nanoelectronic architectures. *Appl Phys A* 80(6):1183–1195
184. Soh C, Quate C, Morpurgo C, Marcus C, Kong C, Dai C (1999) Integrated nanotube circuits: controlled growth and ohmic contacting of single-walled carbon nanotubes. *Appl Phys Lett* 75(5):627–629
185. Sotiriadis PP (2006) Information capacity of nanowire crossbar switching networks. *IEEE Trans Inf Theory* 52(7):3019–3032
186. Spagocci S, Fountain T (1999) Fault rates in nanochip devices. *Proc Electrochem Soc* 98-19:582–596
187. Spielman DA (1996) Highly fault-tolerant parallel computation. In: *Proceedings of the 37th IEEE Symposium on Foundations of Computer Science (FOCS)*, Burlington, 14–16 Oct 1996, pp 154–163
188. Srivastava N, Banerjee K (2004) Interconnect challenges for nanoscale electronic circuits. *TMS J Mater (JOM)* 56(10):30–31
189. Stan MR, Franzon PD, Goldstein SC, Lach JC, Ziegler MM (2003) Molecular electronics: from devices and interconnect to circuits and architecture. *Proc IEEE* 91(11):1940–1957
190. Strukov DB, Likharev KK (2005) CMOL FPGA: a reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices. *Nanotechnology* 16(6):888–900
191. Taubin A, Cortadella J, Lavagno L, Kondratyev A, Peeters A (2007) Design automation of real life asynchronous devices and systems. *Found Trends Electron Des Autom* 2(1):1–133
192. Theis TN (2000) The future of interconnection technology. *IBM J Res Dev* 44(3):379–390
193. Toffoli T (1984) Comment on: Dissipation in computation. *Phys Rev Lett* 53(12):1204
194. Tougaw PD, Lent CS (1994) Logical devices implemented using quantum cellular-automata. *J Appl Phys* 75:1818–1825
195. Tour JM, Van Zandt L, Husband CP, Husband SM, Wilson LS, Franzon PD, Nackashi DP (2002) Nanocell logic gates for molecular computing. *IEEE Trans Nanotechnol* 1(2):100–109
196. Tour JM, Cheng L, Nackashi DP, Yao Y, Flatt AK, St Angelo SK, Mallouk TE, Franzon PD (2003) Nanocell electronic memories. *J Am Chem Soc* 125(43):13279–13283
197. Türel Ö, Lee JH, Ma X, Likharev K (2005) Architectures for nanoelectronic implementation of artificial neural networks: new results. *Neurocomputing* 64:271–283
198. Uchida K (2003) Single-electron devices for logic applications. In: *Nanoelectronics and Information Technology*. Wiley, Berlin, pp 425–443
199. Unger SH (1969) *Asynchronous Sequential Switching Circuits*. Wiley, New York

200. von Hippel AR (1956) Molecular engineering. *Science* 123(3191):315–317
201. von Neumann J (1956) Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. In: *Automata Studies*. Princeton University Press, Princeton, pp 43–98
202. Waingold E, Taylor M, Srikrishna D, Sarkar V, Lee W, Lee V, Kim J, Frank M, Finch P, Barua R, Babb J, Amarasinghe S, Agarwal A (1997) Baring it all to software: Raw machines. *Computer* 30(9):86–93
203. Wang KL, Khitun A, Flood AH (2005) Interconnects for nanoelectronics. In: *Proc. 2005 IEEE International Interconnect Technology Conference*, San Francisco, 6–8 June 2005, pp 231–233
204. Wang W (1990) An asynchronous two-dimensional self-correcting cellular automaton. PhD thesis, Boston University, Boston, MA 02215, short version: In *Proc. 32nd IEEE Symposium on the Foundations of Computer Science*, San Juan, 1–4 Oct 1990. IEEE Press, pp 188–192, 1991
205. Weeber JC, González MU, Baudrion AL, Dereux A (2005) Surface plasmon routing along right angle bent metal strips. *Appl Phys Lett* 87(22):221101
206. Whitesides GM, Grzybowski B (2002) Self-assembly at all scales. *Science* 295(5564):2418–2421
207. Mac Williams FJ, Sloane NJA (1978) *The Theory of Error-Correcting Codes*. North-Holland, Amsterdam
208. Williams RS, Kuekes PJ (2001) Demultiplexer for a molecular wire crossbar network. US Patent 6 256 767
209. Winfree E, Liu F, Wenzler LA, Seeman NC (1998) Design and self-assembly of two-dimensional DNA crystals. *Nature* 394(6693):539–544
210. Wolf SA, Awschalom DD, Buhrman RA, Daughton JM, von Molnar S, Roukes ML, Chtchelkanova AY, Treger DM (2001) Spintronics: a spin-based electronics vision for the future. *Science* 294(5546):1488–1495
211. Wong HSP, Frank DJ, Solomon PM, Wann CHJ, Wesler JJ (1999) Nanoscale CMOS. *Proc IEEE* 87(4):537–570
212. Wood J, Edwards TC, Lipa S (Nov 2001) Rotary traveling-wave oscillator arrays: a new clock technology. *IEEE J Solid-State Circ* 36(11):1654–1665
213. Worschech L, Beuscher F, Forchel A (1999) Quantized conductance in up to 20  $\mu\text{m}$  long shallow etched GaAs/AlGaAs quantum wires. *Appl Phys Lett* 75(4):578–580
214. Wu W, Jung GY, Olynick DL, Straznicki J, Li Z, Li X, Ohlberg DAA, Chen Y, Wang SY, Liddle JA, Tong WM, Williams RS (2005) One-kilobit cross-bar molecular memory circuits at 30-nm half-pitch fabricated by nanoimprint lithography. *Appl Phys A* 80(6):1173–1178
215. Yamada T, Akazawa M, Asai T, Amemiya Y (2001) Boltzmann machine neural network devices using single-electron tunneling. *Nanotechnology* 12(1):60–67
216. Yanagida T, Ueda M, Murata T, Esaki S, Ishii Y (2007) Brownian motion, fluctuation and life. *Biosystems* 88(3):228–242
217. Yang T, Kiehl R, Chua L (2001) Tunneling phase logic cellular nonlinear networks. *Int J Bifurc Chaos* 11(12):2895–2911
218. Zhirnov VV, Cavin RK, Hutchby JA, Bourianoff GI (2003) Limits to binary logic switch scaling – a gedanken model. *Proc IEEE* 91(11):1934–1939
219. Zhong Z, Wang D, Cui Y, Bockrath MW, Lieber CM (2003) Nanowire crossbar arrays as address decoders for integrated nanosystems. *Science* 302(5649):1377–1379

## Nanoscale Atomic Clusters, Complexity of

ANATOLY I. FRENKEL<sup>1</sup>, JUDITH C. YANG<sup>2</sup>,  
DUANE D. JOHNSON<sup>3</sup>, RALPH G. NUZZO<sup>4</sup>

<sup>1</sup> Department of Physics, Yeshiva University,  
New York, USA

<sup>2</sup> Department of Mechanical Engineering and Materials  
Science, University of Pittsburgh, Pittsburgh, USA

<sup>3</sup> Department of Materials Science and Engineering,  
University of Illinois at Urbana-Champaign,  
Urbana, USA

<sup>4</sup> Department of Chemistry, University of Illinois  
at Urbana-Champaign, Urbana, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Synthesis of Nanoparticles Through Self-Assembly  
Observation and Characterization:

Synergy of Experiments and First Principle Modeling

Structural Relaxation of Nanoparticles

Future Directions

Acknowledgments

Bibliography

### Glossary

**Nanoparticles** These, also called nanoclusters, are atomic agglomerates having the size less than 100 nm at least in one dimension.

**Bulk material** The material which grain sizes are in the micrometer scale or larger.

**Self-assembly method** A method that a system uses to develop and form particular structure using only pre-existing components.

### Definition of the Subject

Freestanding and supported, nanoscale metallic clusters (i. e., nanoclusters or nanoparticles) constitute complex systems as they span an enormous space of potential structures, each having different collective properties, with only few potentially stable (or metastable) structures that exhibit technologically beneficial properties. The synthesis and design of collective properties is a critical area of current research – with technological beneficial outcomes often equivalent to finding a “needle-in-the-haystack”. Yet

the synthesis and application of nanoparticles with specific properties has had, and will have, an increasing impact on technology, whether for improving toothpaste, 'green(er)' energy production (e.g., in petroleum refinement, catalysis, or batteries), biotechnology and national security issues (e.g., biosensors). From the perspective of theory, freestanding or supported nanoparticles constitute an enormous challenge in terms of property assessment as well as comparison of results, where the experimental processing routes to produce samples must be considered within theory for one-to-one comparison to measured data and understanding of the controlling physics to that observed.

Here we provide a glimpse into the complexity of the search space and a subset of the critical experimental and theoretical analysis tools that are capable of providing quantitative assessment of structure-property relations actually found, such as assessment of structure and reactivity, say, for catalytic performance. Hence, it is necessary (and desirable) to have a direct and meaningful coupling of experimental and theoretical analysis. It is our purpose to provide a brief overview of current quantitative methods and examples of such coupling of experiment and theoretical analysis.

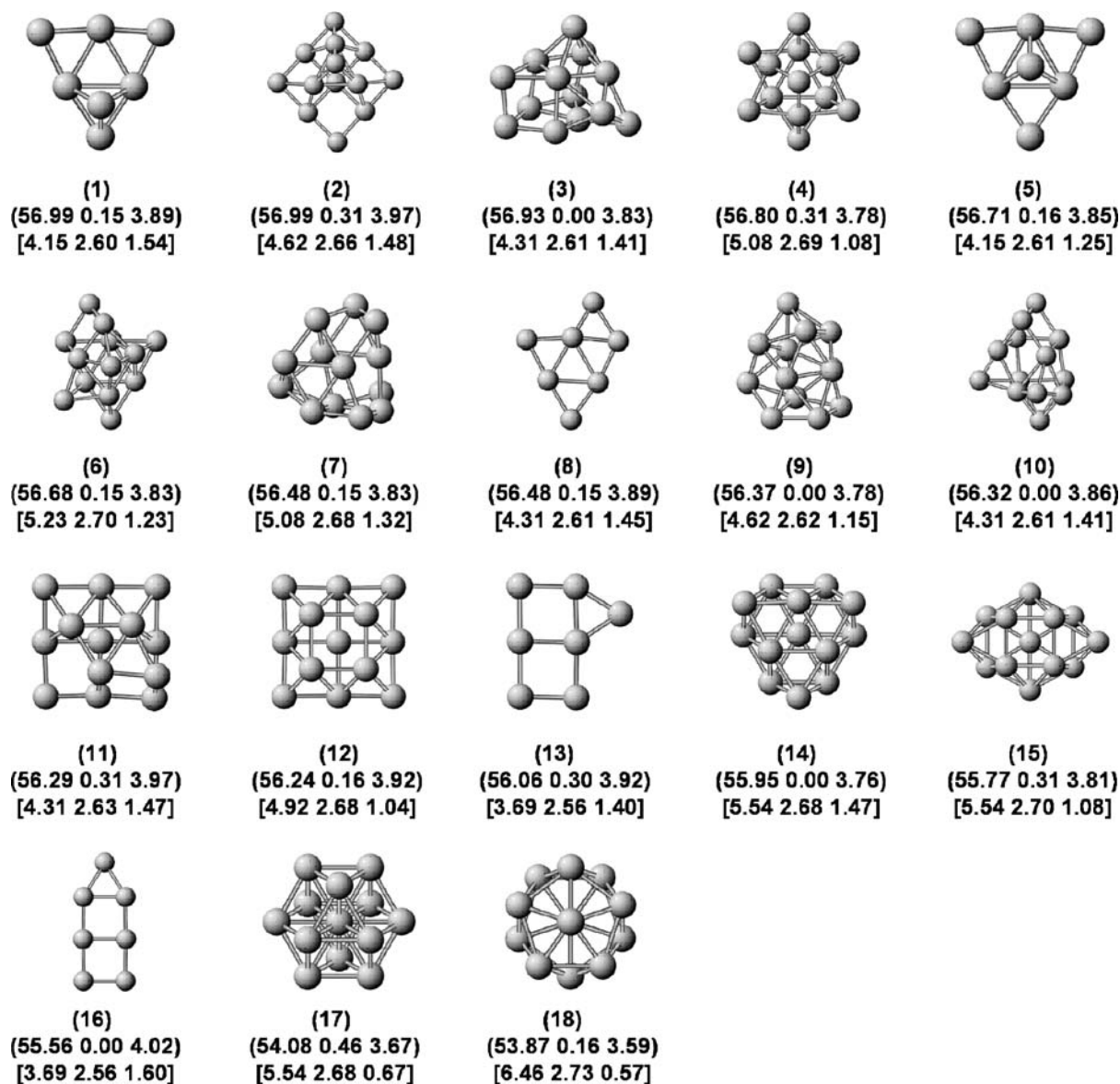
First, to exemplify the challenges, we provide an example list of issues. Even with an established synthesis route for a single-component nanoparticle, one cannot control with certainty the size, only the distribution – a narrow distribution is great but the nanoparticles properties will still depend, at a minimum, on surface-to-volume ratio (i.e., inversely as particle radii or size,  $1/R$ ), an example of which is melting [1]. Already with a two-component nanoparticle, such as Pb-Bi, melting is controlled again in part by  $1/R$  but also by thermal and electronically-driven effects such as chemical segregation, leading, for example, to shell-like nanoparticles with differing melting temperatures for inner and outer regions [2]. In addition, for a particle of fixed number of atoms,  $N$ , there can be an enormous number of structural isomers possible, or bonding motifs. For example, in Fig. 1 are the calculated lowest-energy, 13-atom ( $N = 13$ ) freestanding isomers of Pt, [3] similar results have been obtained for other elements. From Refs. [3,4,5] it can be garnered (amongst many other points) that (i) high-symmetry structures are not necessarily low in energy, (ii) structures can be degenerate in energy, (iii) there is no simple correlation between structure and local properties, such as coordination number or bond lengths, (iv) both 3d and 2d structures compete in energy, and a structure (and concomitantly their properties) will be strongly affected by placing the particles on a support, and (v) intuition regarding

the important structures can be helpful but not reliable. Additionally, for fixed  $N$  and only two types of atoms (A and B) within the cluster  $A_m B_n$ , such that  $N = m + n$ , there are a binomial number  $N!/(m!n!)$  of possible "alloy" configurations (i.e., homotopic groups of clusters, homomers that include stereoisomers) within a single structure (The term "homotops" has been introduced by J. Jellinek to describe such clusters that form homotopic groups. But, to be consistent with standard chemistry use of isomers and the general symmetry descriptions, such as enantiomers, diastereomers, and stereoisomers, we refer to these as homomers). There have been two recent and excellent reviews [6,7] thoroughly discussing these issues, mainly from a theoretical perspective. Finally, nanoparticles with structures that have useful properties may be only possible through self-assembly of organic stabilizers, e.g. Au nanoparticle stabilized via thiol ligands [8]. Clearly, the possible configuration space is daunting. Yet, it is critical that we can assess the structural properties of synthesized freestanding and supported nanoparticles in order to sort out many factors influencing their complex behaviors and correlate them with property design requirements. This is the principal scope of what follows, where we focus on the integrated use of several advanced analytic techniques – x-ray absorption fine structure (XAFS), and atom-counting methods of electron microscopy, and ab initio theoretical methods to achieve these goals and answer questions central for many complex systems, including nanoscale atomic clusters.

## Introduction

Metallic nano-sized clusters (i.e., nanoclusters or nanoparticles) play a crucial role in modern science and technology. The application of design at the atomic scale can be found in petroleum refinement, catalytic converters, and nanofabrication. On a more academic level, metallic nanoparticle research can be found integrated in materials science, biotechnology, and organic chemistry [8]. Despite the ever-increasing interest of nanoparticles in different fields of study, catalysis remains perhaps the essential application of metallic nanoclusters, due to its potential technological impact.

The enhanced catalytic activity of these nanomaterials results from an increased number of atoms exposed to the surface compared to the corresponding bulk material. As a corollary, the number of neighbors to which a given atom is coordinated decreases, giving rise to nonbulk-like behaviors [9,10,11,12,13]. The energetic considerations of such systems lead to important synthetic pathways such as hydrogenation of olefins and silylation of polymers [14].



#### Nanoscale Atomic Clusters, Complexity of, Figure 1

Pt<sub>13</sub> isomers in order of decreasing binding energy calculated via density functional theory (see text). Properties listed below the structure are (in parentheses) energy (eV), magnetic moment ( $\mu_B/\text{atom}$ ), mean inter-atomic distance ( $\text{\AA}$ ), and [in brackets] first nearest-neighbor coordination number, first NN bond length ( $\text{\AA}$ ) and s-d hybridization index, see Ref. [3]. Structures 17 and 18 are the compact O<sub>h</sub> and I<sub>h</sub> isomers, and 14 is the 2-layer cuboctahedrally stacked cluster, which forms the basis of bulk fcc Pt. Reprinted figure with permission from [3]

The syntheses of these nanoscale materials can be categorized as physical or chemical methods. Physical methods involve a “top-down” approach where starting material is sculpted down to the atomic scale in the gaseous state. Once in the gaseous state the atoms can then condense into larger structures. Examples of physical methods in the synthesis of nanoclusters are metal-vapor depo-

sition [15,16,17] and laser ablation [15,18]. In both these methods different sources of energy are utilized to generate atoms from bulk material. Typically, the vaporized atoms are then deposited onto a substrate under vacuum, where particles form by nucleation [19]. Again, the reactive surface of these particles results in aggregation and reduction of the surface area leading to self-assem-



bly of a metallic cluster [20]. One of the drawbacks associated with this method of synthesis is the financial burden due to the expensive equipment required. Another limitation is the broad distribution of cluster sizes generated in the syntheses [15,18]. Yet despite recent advancements in physical methods for nanoclusters synthesis, [21,22,23] the field is predominantly controlled by chemistry [15].

Chemical methods take a “bottom-up” approach toward synthesis. Unlike physical methods, applying chemistry toward nanoparticle synthesis implies the assembly of the structures from chemical precursors. Traditionally, nanoparticle synthesis involves the reduction of metal salts in solution followed by the use of a capping agent to prevent aggregation of the particles. Reducing agents such as alcohols, [24] hydrogen gas, [25] and hydrides [26] can be used in the reduction of the ionic metal. Capping agents typically used are organic polymers, [24] as well as surfactants [25,26]. Coupling of the reducing and capping agent can be accomplished with molecules such as sodium citrate [27]. Chemical methods offer better control of particle size, with the assistance of capping agents, and greater flexibility because of the variety of chemical precursors that can be utilized in synthesis. In contrast to physical methods, fabrication of nanoparticles using chemistry is relatively inexpensive. Even with the improved precision intrinsic to chemical methods however, reproducibility of size, structure, and catalytic ability at the nanoscale regime (1–10 nm) is not concrete. On the other hand, developments in chemical synthesis of nanoparticles have opened doors toward other innovative methods of generating metallic nanoclusters [14].

Despite the relatively large number of available structural techniques, most of them obtain the overall, volume-average properties of nanoparticles, and give little, if any, insight into sometimes very elaborate *actual* arrangement of atoms within the particle. Two state of the art methodologies, synchrotron x-ray absorption fine-structure (XAFS) and quantitative scanning transmission electron microscopy (STEM) have been best positioned for determining the 3D structure and structural habits, both individually and as an ensemble, critical for understanding metallic nanoclusters. XAFS technique is one of the premiere tools to study both atomic and electronic structure of small ensembles due, in part, to its local structure sensitivity and excellent spatial resolution. By measuring coordination numbers, bond lengths and their disorder up to the 5th coordination shell, one can reliably determine the size of the nanoparticles, their shapes (e. g., oblate, raft-like, or truncated polyhedral), surface morphology as well as effects of surface disorder in 1–2 nm-size clusters as

a function of external conditions (temperature, alloy composition, support material, etc.).

Complementary information on site-specific structure and chemistry can be obtained by STEM, which has a unique capability for providing structural and spectral information simultaneously. Supported metal nanoclusters of size of 1–100 atoms make an ideal system for examination by high-angle annular dark-field (HAADF, also known as *Z*-contrast) imaging. Correlating the absolute image intensity to the scattering cross-section has been advanced recently within the STEM-based imaging method. With this improvement, one can directly count, with accuracy of  $\pm 2$  atoms, the number of atoms in a supported nanocluster avoiding complexities associated with coherent diffraction. By utilizing state of the art electron and x-ray probe methodologies, one can explore substrate/nanoparticle interactions as a function of support and nanoparticle material, as well as by size, composition and 3-D structure of the supported nanoparticles.

Most state-of-the-art experimental work is integrated with theoretical calculations to help interpret and accelerate identification of cluster bonding motifs, possible metastable structures, and determine electronic properties and reactivity of relevant clusters. Realistic nanometer-sized, metallic clusters (i. e., those equivalent to experiment) can be reliably simulated by electronic-structure and molecular-dynamic techniques to address the issues of complex geometries (cluster size and corresponding thermodynamically- or kinetically-stabilized shapes and atomic arrangement), of shape evolution (kinetics), as well as local bonding effects that determine reactivity. As such, a correlated, self-consistent interpretation of the experimental data from methods discussed above can accelerate identification of cluster bonding motifs, possible metastable and dynamic structures, and determination of the electronic properties of relevant clusters.

In the remainder of the article will first review the methods of synthesis of nanoparticles, including those guided by self-assembly (Sect. “[Synthesis of Nanoparticles Through Self-Assembly](#)”). The modern experimental and theoretical methods of nanoparticle analysis will be reviewed in Sect. “[Observation and Characterization: Synergy of Experiments and First Principle Modeling](#)”. Complexity of several typical cluster systems, mono- and bimetallic, supported and freestanding, will be illustrated by examples from recent works (Sect. “[Observation and Characterization: Synergy of Experiments and First Principle Modeling](#)”). Sect. “[Structural Relaxation of Nanoparticles](#)” discusses the possible origins of structural relaxation and surface reconstruction observed experimentally



and studied theoretically in ultra-small clusters. Future directions are outlined in the final section.

### Synthesis of Nanoparticles Through Self-Assembly

Besides traditional methods of chemical synthesis, the field has expanded to include many other routes toward production of metallic nanoparticles. Among these newer techniques are photochemical synthesis, electrochemical synthesis and sonochemical synthesis. Another departure from standard chemical methods of nanoparticle synthesis is in the production of bimetallic nanoparticles [7]. The classic framework for creating nanoparticles, although efficient for some elements, requires various reagents and consequently rigorous reaction conditions [28].

Photochemically induced synthesis of nanoparticles offers the advantage of using ultraviolet light to assist in the reaction without the addition of capping or stabilizing agents [14,28,29]. The reduction reaction can be explained by the photoexcitation of a metal salt and organic molecules containing carbonyl groups. The photoexcitation of the organic compound often results in the formation of radical [28,30] or subsequent formation of a radical [29]. Reduction of the metal ion ensues by the formed organic radical. In the case of multivalent metals, reduction occurs until the atom reaches a neutral state and aggregation of a nanoparticle follows thereafter. Although the formed particles aggregate without the inclusion of a capping agent, crystal growth is mediated by the ionic liquid solution of the original reaction mixture. The surfaces are stabilized electrostatically by the concentration of ions in the solution [28,29]. Other methods of photochemical synthesis often use reducing agents which also serve as capping agents; physically adsorbing to nanoparticles and arresting further aggregation [30,31,32,33].

Another alternative to traditional methods of synthesis is sonochemistry. The inner workings of sonochemistry stem from the extreme temperatures and pressures achieved by the expansion and implosion of bubbles in solutions, a process known as cavitation [14,34,35]. The intense energy released upon collapse of the bubbles results in the decomposition of molecules and the formation of free radicals. As with photochemical synthesis, the resulting free radicals initiate the reduction reactions used to reduce the metal ions in solution [36]. Particle size is maintained by the use of capping agents [37,38,39]. The main advantages of using sonochemical methods are the fast reaction rates and the ability to generate very small sized nanoparticles. On the other hand, a broad distribution of particle sizes is also typically obtained from this type of synthesis [40].

Electrochemical synthesis of nanoparticles is another viable option capable of generating metallic nanoclusters. In this method an anode is placed in solution and oxidized in the presence of a chaperone molecules or stabilizing agents. The oxidized atoms from the anode migrate to the cathode where they are reduced. Aggregation shortly follows after the migration to the cathode where cluster growth is mediated by the stabilizing agents. Isolation of the particles is accomplished by precipitation of the product from the cathode [14,41,42,43]. There are methods where a rotating cathode [44] or a double pulse technique [45,46] is implemented to improve size distribution and yield, but ultimately the chemistry that occurs is identical to the standard electrochemical method. Comparatively, reduction using electrochemistry has many advantages over its counterparts. Once formed, the nanoparticles can be easily isolated as they begin to precipitate out of solution. The size of the clusters generated can be controlled varying the current intensity of the cell, [42,45,46] but perhaps the greatest advantage of this method is the high yields achieved [14].

Another area of synthesis important in the field of metallic nanoparticles is bimetallic nanocluster preparation. The interest in bimetallic nanoclusters arises from the changes in physical and chemical properties compared to monometallic species [47,48,49,50,51,52]. As with monometallic clusters, many routes exist toward the synthesis of bimetallic clusters. These materials can be synthesized by traditional forms of reduction, using metal ion-containing salts (precursors). Additionally, this system of reduction can be accomplished simultaneously or sequentially. In simultaneous synthesis both metallic precursors are reacted in the same solution in the presence of capping agents. The mixture can result in the spontaneous formation of a core/shell structure as well as other alloyed structures [39,53]. Alternatively, successive addition of the metallic precursors has proven to be a more efficient way to control particle structure [14,39,54,55]. In the successive method the formation of the initial metal cluster behaves as a nucleation site for the growth of the second metal, affording a core shell motif. Incidentally, this self-assembly of core/shell structures has been induced simply by the mixing of nanoclusters of two different metals [56,57]. The structural motif has been attributed to a balance between surface energies as well as binding energies where size and composition of the clusters is detrimental to structural refinement [56]. Studies have also shown that the core/shell structure can be inverted at extreme temperatures [53]. Furthermore bimetallic nanoclusters can be synthesized by using a sacrificial layer of hydrogen adsorbed onto the surface of the core metallic

cluster. Using the adsorbed hydrogen layer as a reducing agent, incoming metallic ions are reduced onto the surface of the core cluster forming a shell. Thus far this method has only been generated with specific metallic species [58]. In addition to the aforementioned methods of synthesizing bimetallic particles, synthesis can also be accomplished by sonochemical, [39] electrochemical [54] and photochemical [55] methods.

### Observation and Characterization:

#### Synergy of Experiments and First Principle Modeling

Combination of complementary structural techniques – electron microscopy and XAFS, as well as DFT/MD simulations, provide previously unavailable atomic-level understandings of the structural dynamics of the most important forms of supported metal clusters. These methods can provide understandings of the metal framework bonding present in supported nanoscale clusters and the significant impacts on them that can originate as a consequence of adsorbate bonding and (more recently revealed) electronic effects mediated by support interactions. The data from microscopy now allows one to establish precise atomic compositions in supported catalyst systems, counting atoms explicitly at the single nanoparticle level. Experimental protocols based on synchrotron x-ray absorption spectroscopy allow the elucidation of the precise structural habits adopted by supported forms of metallic nanoparticles – establishing methodologies of modeling that define the nature of the strains present in such systems; unraveling the complex nature of the atomic level bonding habits they embed; the dynamical factors that mediate transformations of these structures due to the impacts of particle size, temperature-dependent support interactions, and adsorbates; the nature of the size-dependent atomic-bond relaxations that these particles embed; and the nature of the complex electronic structures that are unique to materials of this sort [59,60,61,62,63,64].

Development and testing of these advanced methods rely on the availability of synthetic methods that have provided model systems – both in the form of discrete monolayer protected clusters and supported metal nanoparticle catalysts – with precisely defined compositions and (for the latter) extremely narrow distributions of atomic mass.

Metal nanoclusters [65] (and gold nanoclusters in particular [66]) can exhibit structures that differ significantly from that corresponding to the bulk, ones that depend strongly on cluster size. Detailed knowledge of these structures is crucial for understanding and predicting nanocluster properties, including chemical, electrical, magnetic, and optical ones. The experimental determination

of atomistic structural information is a very difficult task: analyses by imaging or scattering methods are presently limited by insufficient spatial resolution or by the coherent scattering size of these techniques. As a result, most structural determinations proceed indirectly by comparing experimentally accessible properties (e.g., ion mobility, photoemission spectra, polarizability, optical absorption, etc.) with those computed theoretically for candidate structures. Synergistic approaches, however, combining several structural techniques into a self-consistent structure refinement method, make accessible the previously unknown physical picture of the nanoscale. In the following sections, we will show examples of the determination of size, shape and atomic structures of mono- and bimetallic nanoparticles by combining results of extended x-ray absorption fine-structure (EXAFS), advanced methods of electron microscopy and DFT/MD calculations.

#### Structure Determination by EXAFS

By using EXAFS technique, one can extract accurate information about the identities, average distances, and coordinations of the neighboring atoms to the x-ray absorbing atom in nanoclusters [67,68]. The EXAFS signal,  $\chi(k)$ , is the sum of all contributions,  $\chi_i(k)$ , from groups of neighbors at approximately equal distances from the absorbing atoms (i.e., the  $i$ th shell), which are often written as: [69]

$$\chi_i(k) = \frac{S_0^2 n_i}{k R_i^2} \left| f_i^{\text{eff}}(k) \right| \cdot \sin \left[ 2k R_i - \frac{4}{3} \sigma_i^{(3)} k^3 + \delta_i(k) \right] e^{-2\sigma_i^2 k^3} e^{-2R_i/\lambda_i(k)}, \quad (1)$$

where  $k$  is the photoelectron wave number,  $f_i^{\text{eff}}(k)$  and  $\delta_i(k)$  are the photoelectron scattering-path amplitude and phase, respectively,  $S_0^2$  is the passive electron reduction factor,  $n_i$  is the degeneracy of the scattering path,  $R_i$  is the effective half-path-length (which is equal to the interatomic distance for single-scattering paths),  $\sigma_i^2$  is the mean-square deviation in  $R_i$ ,  $\sigma_i^{(3)}$  is the third cumulant, and  $\lambda_i(k)$  is the photoelectron mean free path. Using modern computer packages, e.g., IFEFFIT, [70] which employs a non-linear least square method to fit theoretically calculated (with the help of FEFF6 code [69]) EXAFS signal to the data, one can obtain the best-fit values of structural parameters, together with their uncertainties. The mean-square deviation,  $\sigma^2$ , of the first nearest neighbor (1NN) distance can be represented to a good approximation as a superposition of static ( $\sigma_s^2$ ) and dynamic ( $\sigma_d^2$ ) terms:

$$\sigma^2 = \langle (r - \langle r \rangle)^2 \rangle = \sigma_s^2 + \sigma_d^2. \quad (2)$$

To separate the temperature-independent  $\sigma_s^2$  and temperature-dependent  $\sigma_d^2$ , one can use a simple correlated Einstein model for  $\sigma_d^2$ :

$$\sigma_d^2 = \frac{\hbar}{2\omega\mu} \frac{1 + \exp(-\Theta_E/T)}{1 - \exp(-\Theta_E/T)}, \quad (3)$$

where  $\omega$  is a bond vibration frequency,  $\mu$  is the reduced mass of the 1NN atomic pair, and  $\Theta_E = \hbar\omega/k_B$  is the Einstein temperature. Thus, the total  $\sigma^2$  in this approximation depends on three parameters:  $T$ ,  $\Theta_E$ , and  $\sigma_s^2$ . By replacing the total  $\sigma^2$  in Eq. (2) by a sum of the dynamic and static terms (Eq. (2)), the best fit results for  $\Theta_E$  and  $\sigma_s^2$  can be obtained from a concurrent non-linear least square fitting of Eq. (2) to the EXAFS data taken at various temperatures. From obtaining the best fit values, with their uncertainties, of  $n_i$ ,  $R_i$ ,  $\sigma_s^2$ ,  $\sigma_d^2$  and  $\sigma_i^{(3)}$ , one can significantly reduce the number of possible models of cluster structure and its interaction with the environment (thermal effects, gas atmosphere effects, cluster-substrate, cluster-adsorbate interactions, etc.) EXAFS, however, only yields ensemble-averaged information; thus its application is most valuable for structure refinement of size-controlled nanoclusters.

There are several approaches that allow one to estimate cluster size from EXAFS results. The most popular method is using the first nearest neighbor (1NN) coordination number  $n_1$  that can be directly obtained in EXAFS analysis. These data can be compared against *exact* cluster models or approximate expressions. For the first approach, calculations of Montejano-Carrizales et al. [71] are useful as they obtained geometrical characteristic of several regular polyhedral clusters (cubo-octahedral, icosahedral, body-centered cubic and simple cubic) analytically as a function of the cluster order  $L$ . Defining  $L = N_E - 1$ , where  $N_E$  is the number of atoms along the edge of a regular polyhedron, the following relationships can be derived for the 1NN coordination numbers in cuboctahedral (closed packed) and icosahedral (non-closed packed) clusters that have the same sequence of their magic numbers ( $N = 13, 55, 147, 309, 561, 923 \dots$ ): [71]

$$\begin{aligned} n_1^{\text{co}} &= \frac{24L(5L^2 + 3L + 1)}{10L^3 + 15L^2 + 11L + 3}, \\ n_1^{\text{icos}} &= \frac{6L(20L^2 + 15L + 7)}{10L^3 + 15L^2 + 11L + 3}. \end{aligned} \quad (4)$$

Calvin et al. [72] used the second approach by approximating the cluster shape as a sphere with radius  $R$  and obtaining the 1NN coordination number for a cluster with average 1NN distance  $r$  as follows:

$$n_1 \approx \left[ 1 - \frac{3}{4} \left( \frac{r}{R} \right) + \frac{1}{16} \left( \frac{r}{R} \right)^3 \right] n_1^{\text{bulk}}.$$

This approach is advantageous for larger clusters, and when the size distribution is relatively broad.

The method of measuring of 1NN coordination numbers for size determination in supported metal particles dates back to the end of 1970s [73,74,75,76,77,78,79,80,81,82,83,84]. An early publication detailing an analysis of the structures of supported Pt nanoparticles by EXAFS, found a strong correlation between the measured first nearest neighbor (1NN) metal coordination number and the disorder of their bond lengths; lower coordination numbers appeared to correlate strongly with increased measured disorder in the first-shell metal-metal bond lengths [73]. These effects would be most sensitively seen in particles of the smallest size. The importance of anharmonic corrections to the 1NN pair interaction potential had yet to be appreciated in this pioneering work. Neglecting this effect leads to a non-physical decrease in 1NN bond lengths measured at high temperatures. Enhanced disorder in the 1NN bond distance was later correlated with the influences that result from cluster-support interactions [79]. These studies provide an important insight into the nature of supported metal nanoparticles, namely that the bond lengths of surface atoms should exhibit enhanced structural disorder.

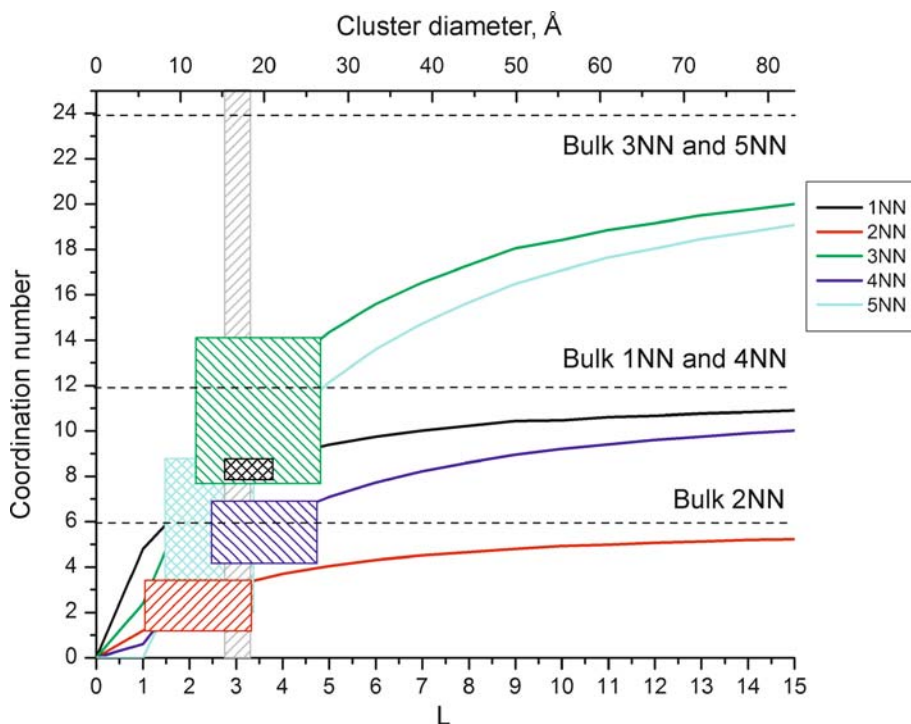
Another approach, to use size-dependent 1NN distance change as a measure of cluster size, was used by Frenkel et al. [61] to obtain sizes of alkanethiolate-stabilized Au nanoparticles. Such changes were first observed by Mays et al. [85] and attributed to surface tension (ST) in the framework of liquid drop model of nanoclusters. To analytically estimate particle diameter,  $d$ , the ST method utilized the equation:

$$d = \frac{4}{3} f_{\text{rr}} K \alpha,$$

where  $f_{\text{rr}}$  and  $K$  are the surface stress and compressibility in the bulk, and  $\alpha = \Delta R/R$  is the relative lattice contraction that can be measured by EXAFS. Montano et al. [86,87,88] studied Fe, Cr, Ag and Cu nanoclusters by EXAFS and observed contractions of the average interatomic distances relative to the bulk in all cases.

An alternative approach by Jiang et al. [89] is based on the surface multiplayer relaxation phenomenon, proposed by Finnis and Heine [90]. According to their model, cluster interatomic distances can be calculated for an arbitrary polyhedral shape, using known experimental relaxation data for macroscopic crystal surfaces.

The 1st shell methods described above are based on analytical calculations of the coordination numbers or distances as a function of particle size. However, when clusters are asymmetric, these methods are not accurate since



**Nanoscale Atomic Clusters, Complexity of, Figure 2**

Cluster size, shape and structure determination: comparison of the average coordination numbers (up to 5NN), together with their error bars, measured by EXAFS for the carbon-supported Pt nanoparticle sample and those predicted from the truncated cuboctahedron model for various cluster sizes. Reprinted with permission from [68]

model values of  $n_1$  may be similar when not only the size, but the shape and structure of the cluster are allowed to vary. Only when coordination numbers (or path degeneracies) corresponding to the more distant shells and multiple-scattering contributions are measured, one will have a series of indices:  $\{n_i\}$ ,  $i = 1, 2, 3 \dots$  which should be unique for any given polyhedral cluster model. Since multiple-scattering contribution to EXAFS in nanoclusters can be quantitatively analyzed, [91] not only the single-scattering coordination numbers but also multiple-scattering path degeneracies could be reliably extracted from the EXAFS data. Such additional information effectively communicates size, shape and surface orientation of nanoclusters (Fig. 2) [68].

Even though the methods above can be used to describe metal-metal coordination numbers and, therefore, size, shape and structure information in monometallic and heterometallic clusters, the latter are characterized by significantly higher level of complexity, due to their chemical heterogeneity.

In bimetallic clusters, two distinctly different types of mixing of A and B atoms are possible. They can be mixed statistically (i. e., randomly, in accordance with the

overall concentration) or non-statistically. The most common example of non-statistical mixing is segregation of atoms of different elements, forming a core-shell-type particle where larger than the concentration-weighted average number of atoms of one type can be found in the core, and the other type – at the surface of the particle. In heterogeneous samples, where different clusters (A-rich and B-rich) can be formed, the situation may be further complicated [92].

For *random* alloys, the average coordination numbers  $n_{AA}$  and  $n_{AB}$  of A and B atoms relative to A atom are in the same proportion as the bulk concentrations of these elements in the sample:

$$\frac{n_{AA}}{n_{AB}} = \frac{x_A}{x_B}.$$

For alloys with nonzero short range order, the left part may be *larger* or *smaller* than the right part, indicating *positive* or *negative* tendency to clustering, respectively. In the former case, the atoms A and B segregate to different regions of the nanoalloy. In the latter, the A atoms are preferentially coordinated with B (with probability greater than  $x_A/x_B$ ) and vice versa.

We can also introduce a short range order parameter,  $\alpha$ , analogously to its definition by Cowley for bulk alloys: [93]

$$\alpha = 1 - \frac{n_{AB}/n_{AM}}{x_B},$$

where  $n_{AM} = n_{AA} + n_{AB}$  is the coordination number of the A-metal bonds. For alloys with positive or negative tendency to clustering,  $\alpha$  will be positive or negative, respectively. However, even after the segregation is demonstrated by examining the experimental values of  $n_{AA}/n_{AB}$  or  $\alpha$ , more experimental information is still needed to find out whether A is predominantly in the surface or in the core, as well as for the determination of the particle size.

Such information is available by measuring EXAFS on both A and B central atoms and extracting coordination numbers  $n_{AA}$ ,  $n_{AB}$  and  $n_{BB}$ . The analysis should be done concurrently, with obvious constraints imposed on the heterometallic bonds during the fits: [94]

$$n_{AB} = \frac{x_B}{x_A} n_{BA}; \quad R_{AB} = R_{BA}; \quad \sigma_{AB}^2 = \sigma_{BA}^2$$

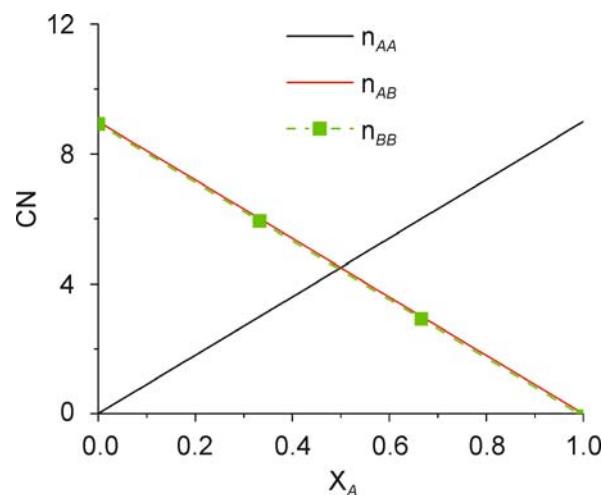
The atoms of the type A will segregate to the surface of the nanoparticle and B – to the core, if  $n_{AM} < n_{BM}$ , since atoms at the surface have fewer neighbors than those in the core. This criterion is useful even for alloys containing elements that are neighbors in periodic table (e.g., Fe-Ni, Pd-Ag, etc.) where only the total  $n_{AM}$ ,  $n_{BM}$  numbers can be measured by EXAFS analysis of A and B absorbing atoms, respectively, due to the similarity of backscattering amplitudes of A-A and A-B pairs (as well as B-A and B-B). Another advantage of analyzing both A and B EXAFS data is for the particle geometry determination. Indeed, the average number of metal-metal neighbors per metal atom:

$$n_{MM} = x_A n_{AM} + x_B n_{BM}$$

for the first nearest neighbor shell, combined with other information (e.g., higher shell coordination numbers, transmission electron microscopy (TEM) data, etc.) allows one to estimate the particle size by methods similar to those described above for monometallic particles.

Random bimetallic alloys have a unique behavior of these coordination numbers with concentration. Assume, for simplicity, a bimetallic nanoparticle of a certain size, with random distribution of A and B atoms, where the following relationships apply:

$$\begin{aligned} n_{AM} &= n_{BM} = n_{MM}; & n_{AA} &= n_{BA} = x_A n_{MM}; \\ n_{BB} &= n_{AB} = x_B n_{MM} &= (1 - x_A) n_{MM}. \end{aligned}$$



Nanoscale Atomic Clusters, Complexity of, Figure 3

Theoretical partial coordination numbers in random nanoalloys (assuming  $n_{MM} = 9$ ) as a function of composition

Thus, partial coordination numbers should depend linearly on alloy concentration (Fig. 3) in random nanoalloys, provided that the particle size is the same at all concentrations.

### Structure Determination by Advanced Methods of Transmission Electron Microscopy

As mentioned above, while multiple-scattering (MS) EXAFS allows determination of the bonding geometry in the nanoscale, it does so in the form of ensemble-average information. Advanced TEM and scanning TEM (STEM) methods, on the other hand, can probe the atomic-level structure and chemistry of individual nanoparticles and allow direct visualization of nano-scale phenomena [59]. Modern electron microscopy is a powerful tool which, if used in conjunction with EXAFS, can provide a wealth of information about the samples which otherwise could not be determined if these techniques were not used synergistically.

Using “Z-contrast” protocol in STEM, one can correlate the absolute image intensity to the scattering cross-section, so that the number of atoms can be counted experimentally. Furthermore, high-resolution electron microscopy (HREM) can visualize the facets and shapes of nanoparticles and the interfacial structures between the nanoparticles and their supports. In this way, a full mapping can be made of the structural landscape present in a complex heterogeneous sample.

Transmission electron microscopy has been used extensively and successfully to study nanoparticles [95]. It



has a higher level of accuracy than other methods for samples containing small number ( $< 100$ ) of atoms [59]. Specifically, TEM methods have been used to examine clustering sites as well as the shapes of various clusters [95]. For example, one specific problem of interest in current heterogeneous catalysis research is determining the elemental distribution in a bimetallic catalyst particle, e.g., homogeneous distribution of elements or core-shell structures with separation of elements to the core and surface of the particles. In addressing these questions, TEM yields a wealth of information. Conventional TEM is capable of direct visualization at the resolution level of  $< 0.3$  nm in standard instruments and  $< 0.1$  nm in aberration-corrected microscopes. Such resolution allows the investigation of structure of catalysts at atomic level. Transmission electron microscopy uses a parallel beam of electrons, whereas analytical TEM, or STEM, methods employ a scanning attachment on the instrument that focuses the electron probe to enable energy dispersive spectroscopy (EDS) and electron energy-loss spectroscopy (EELS) detection to study elemental distribution as well as oxidation states from specific nanostructures in the sample. For example, a VG-STEM is an ultrahigh vacuum (UHV) instrument, and very fine probe ( $\sim 0.5$  nm diameter) with the ability to simultaneously image (Z-contrast) and collect EDS and EELS to acquire both structure and elemental distribution information.

One of the most common TEM techniques is HREM, which is based on phase contrast, and has been used extensively in nanoparticle studies as well. Thomas and Midgley [96] note that the two major techniques in utilizing HREM for atomic and nano-scale structures are the multislice method and basic electron diffraction studies which are analogous to x-ray crystallography. In using the multislice method one essentially matches theoretical quantum mechanical calculations of the electron diffraction patterns of the sample with the closest data set taken in order to infer information on the electronic and crystal structure of the sample [96].

The method works by essentially dividing the reciprocal space into “slices” with an “electron field/wave function” corresponding to each slice. The  $n$ th field of the  $n$ th slice is given by [97]:

$$\phi_{n+1}(b - \Delta b) = [\phi_n(b) \cdot q_n(b)] * \left(\frac{k_z}{z}\right) p(b)$$

where  $b$  is a reciprocal space vector and  $\Delta b$  is the shift in the origin of the wave function and the surface at each slice.  $q_n(b)$  is the “phase grating” and  $p(b)$  is the free space propagator. The asterisk represents the convolution of the two functions [97].

Transmission electron microscopy has an additional advantage over other techniques such as atomic force microscopy (AFM) in that the analysis of x-rays created by the electron beam/sample interaction can be used for elemental mapping [96]. The physical and chemical structure of the sample, e.g., bond structure (or electronic structure) can be obtained via analysis of the EELS. From the energy spectra of scattered or transmitted electrons, one can determine specific processes that occur in the sample, i.e., plasmon resonances and phonon excitations. These in turn offer a way to determine the electronic structures which give rise to such events in the material. The techniques include energy loss near edge structure (ELNES) which contains the chemical bond information. One can also coordinate the various states of excited atoms using the extended energy loss fine structure (EXELFS), a technique similar to EXAFS.

Determining the density of states and electronic structures of material interfaces is of particular importance in catalysis because the electronic structure at the interface governs aspects of adsorption, charge transfer and bonding. Thus EELS can be viewed as complementary to the other techniques mentioned above in allowing a detailed understanding of the chemical properties of individual nanoparticles [95] as well as their interaction with the substrate surface [98]. For example, Liu found that by using EELS analysis on a Pd-Ni bi-metallic system he could determine structural information about the thickness of Pd shell layer [98].

Perhaps the most important set of tools in nanocatalysis studies are incoherent elastic scattering techniques that are usually performed at large scattering angle. In the case of the high-angle scattering signals the incoherent scattering intensity of the signals depends on the atomic number,  $Z$ , as described by Rutherford scattering theory, which is ideal for imaging heterogeneous catalytic materials where high  $Z$  metals, e.g., Pt, are dispersed on a low  $Z$  support, e.g., alumina. Quantitative STEM which is based on high-angle annular dark-field (HAADF), or quantitative Z-contrast imaging, is the technique based on the acquisition of these very high scattering signals ( $> 96$  mrad) [95]. HAADF is especially powerful in catalysis studies because the electronic structure and catalytic activity are directly related to the number of atoms in the particle [95]. Because many particles can be imaged simultaneously by HAADF, then this quantitative STEM method provides better statistical information, such as particle size distributions [95]. A major advantage of STEM over HREM is that the technique readily allows for EELS analysis, and x-ray emission analysis at the nano-level where scattered electrons can be simultaneously collected [96,99].

Z-contrast exploits nonlinear interference effects between the wave-packets of the scattered electrons [99]. The intensity of scattered electrons can be obtained from the expression for Rutherford scattering cross section:

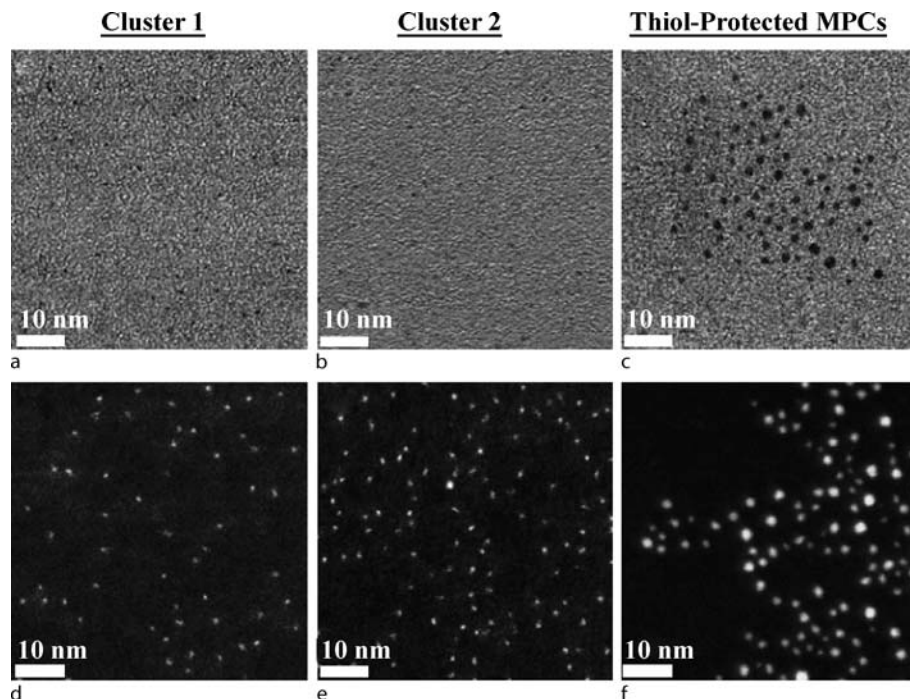
$$\frac{d\sigma}{d\Omega} = \frac{1}{16} \left( \frac{Z^2 e^2}{p^2/2m} \right) \left( \frac{1}{\sin^4(\theta/2)} \right)$$

Since the differential cross-section of the scattered electrons is proportional to  $Z^2$ . A large contrast exists between the supporting structures and the nanoclusters in the image. By collecting electrons scattered to very high angles ( $> 96$  mrad), one can suppress the in-column coherent such that the intensity of the image is simply the number of atoms in the cluster times the intensity expected for one atom of that element. Individual atoms in a particle within the sample may be counted provided that the detector has been properly calibrated. When the calibration is done, one then simply divides the total measured cross section by the cross section of an individual atom [100].

As can be seen in the Fig. 4, the contrast between the standard bright-field (BF) STEM images and the Z-contrast images can be striking. The images are those of monolayer protected gold nanoparticles. There is a large

contrast in the HAADF image where one can distinguish between two elements of greatly varying atomic numbers, namely the higher atomic number gold versus carbon [101]. In that particular study Z contrast was successfully employed to determine particle diameters of the various gold clusters as well as count atoms in individual particles [101].

Since the scattered signal is directly proportional to the number of atoms in the scattering center, one can calculate particle diameters measuring both the intensity profiles of the gold and then subtracting the profile of the substrate/supporting material alone, resulting in that of the gold particles [101]. Intensity profiles across the individual gold particles can then be plotted and measured at full width at half maximum (FWHM). The relationship between the number of atoms in the particle and its diameter is characteristic of the particle's shape. By comparing the relative intensities as a function of particle size, one can infer the nucleation and growth behavior of the particles. In general, the measured intensity and particle growth modes are intimately related [101]. If the intensity is directly proportional to the diameter cubed then it can be inferred that the particles are growing uniformly in three dimensions. However, if the intensity is proportional to the diameter



**Nanoscale Atomic Clusters, Complexity of, Figure 4**

Representative BF-STEM images of a mixed ligand (thiols/phosphines) Cluster 1, b mixed-ligand Cluster 2, and c thiol-protected clusters. Representative HAADF-STEM images of d Cluster 1, e Cluster 2, and f thiol-protected clusters. All images collected at 1 Mx magnification. Parts a and d are images of the same area collected simultaneously. Reprinted with permission from [100]

squared, the particles are seen to be forming a two-dimensional mono-layer or bi-layer [102].

The annular dark-field detector is calibrated by tilting the beam directly onto the HAADF detector, such as by using micro-diffraction mode, with an attenuated electron beam current in order to prevent saturation of the detector. In calibrating the dark-field (DF) detector the current is reduced by a factor of 100. The image of the HAADF detector provides the angular dependence of the detector efficiency, which is taken into account when quantifying the experimental electron scattering [100]. The nanoparticles are then imaged using the very high angle annular dark-field detector such that electrons scattered above  $\sim 100$  mrad contribute to the image intensity. The number of atoms can be quantified as follows:

$$N_0^{\text{high}} = \left( \frac{\sum_{\theta=\beta_{\min}}^{\theta=\beta_{\max}} N_0^{\text{low}}(\theta) f(\theta)^2 2\pi \sin \theta d\theta}{\sum_{\theta=\beta_{\min}}^{\theta=\beta_{\max}} f(\theta)^2 2\pi \sin \theta d\theta} \right) \left( \frac{I_{\text{high}}}{I_{\text{low}}} \right)$$

where  $\beta_{\min}$  is the inner detector angle (130 mrad),  $\beta_{\max}$  is the outer detector angle,  $N_0^{\text{low}}$  is the detector response at the attenuated current,  $f(\theta)$  is the atomic electron scattering factor of the element,  $I_{\text{high}}$  is the current of the electron beam used to image the particles, and  $I_{\text{low}}$  is the attenuated beam current used to image the detector for calibration [100]. Once the calibrations have been performed the intensity of scattered electrons from the nanoparticles can be measured and the background signals can be subtracted. Once the particle intensities are determined the data can be used to count the number of atoms.

Menard et al. demonstrated the utility of these techniques in studying gold nanoparticle monolayer protected clusters on carbon supports using the quantitative HAADF method and EXAFS [59,100,102]. By utilizing these techniques in tandem, characterization of the ligand-protected Au<sub>13</sub> nanoparticle size, shape, and structural disorder was accomplished with an uncertainty of about three atoms. The combination of techniques allows for a detailed analysis of the transitioning point in terms of numbers of atoms, between classical and quantum behaviors or “bulk metallic and molecular states” as well as supporting the theoretical calculations such as those of the scattering amplitudes [59]. In these studies the HAADF imaging was performed with a field-emission Vacuum Generator HB501 STEM at 100 kV. The “atom counting” images were taken at 1-million magnification and have dimensions of 1024 × 1024 pixels.

Further HREM measurements were performed utilizing a field-emission JEM 2010F TEM/STEM operated at 200 kV to confirm the size and observe the crystallinity of the individual nanoparticles [59]. The importance of us-

ing the quantitative Z-contrast and HREM data is that it provides a set of constraints on the interpretation of corresponding EXAFS data as well as determines the grafting density of thiolates on the surface of the particles [59].

This summary demonstrates that high spatial resolution electron microscopies (such as quantitative Z-contrast) and HREM, are extremely useful in determining the behavior of nanomaterials, including nanoparticles, whose properties are transitional between “bulk” and “molecular” materials.

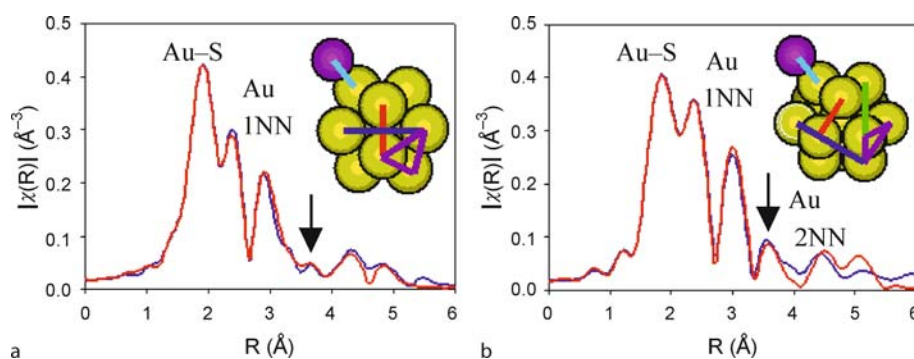
### Cluster Studies by First Principle Theories and Molecular Dynamic Simulations

Theoretical methods also can greatly advance understandings of catalytic materials. First-principles, molecular dynamic (MD) simulations based on density functional theory (DFT) have been advanced to a point now where they can play a significant role in the interpretation of experimental data, revealing critical factors that determine and underlie specific metal cluster properties, see recent reviews [6,7]. (For brevity no attempt is made to include the many pertinent theoretical investigations and new methods, only a few critical papers regarding each key issues.) It immediately becomes more difficult to do this, however, as the cluster size or the chemical complexity (such as due to promoters, defects, adsorbates) of the system is increased, as noted already. In such cases, the DFT search space becomes daunting. Theory then becomes increasingly dependent on the direct experimental input and the availability of computing resources. Even with such capacities, however, the ability to identify relevant low-energy structures lying within a complex, and chemically rich, energy landscape remains a significant challenge and the insights coming from chemical “intuition” remain an important resource for developing useful analytical models.

### Examples of Integrated Methods of Cluster Analysis

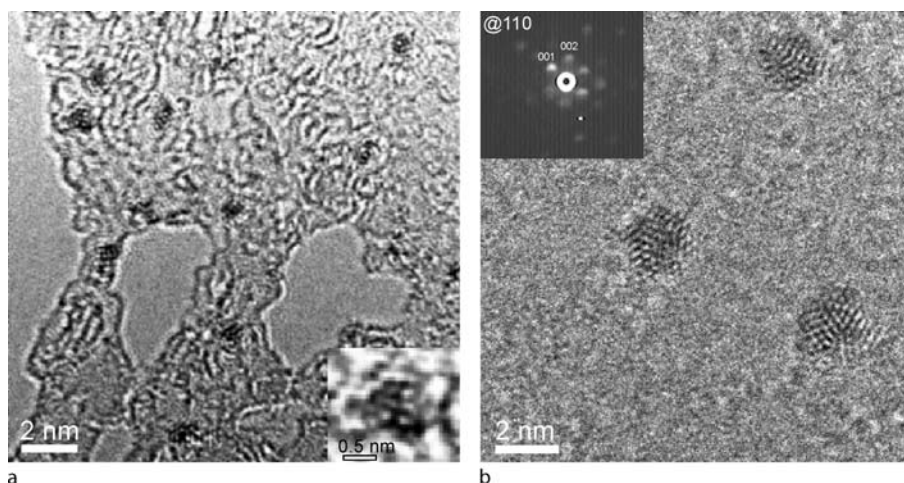
A recent combination of EXAFS and atom counting methods of transmission electron microscopy (TEM) on such specially-synthesized nanoclusters (Au<sub>13</sub>[PPh<sub>3</sub>]<sub>4</sub> [S(CH<sub>2</sub>)<sub>11</sub>CH<sub>3</sub>]<sub>4</sub>) has found them to be highly monodisperse, with their overall structure possessing, on average, 13 gold atoms, with Au-Au coordination number of  $6.7 \pm 0.7$ , an average Au-Au bond length of  $2.85 \pm 0.02$  Å, and an average Au-ligand distance of  $2.324 \pm 0.007$  Å [59]. The presence of eight ligands per cluster was deduced from x-ray photoelectron spectroscopy data [100].

The data presented in Fig. 5a,b illustrate such integrated analyses as applied to several Au nanoclusters – the



#### Nanoscale Atomic Clusters, Complexity of, Figure 5

Fourier transformed EXAFS spectra (blue) and fits (red) for **a**  $\text{Au}_{13}(\text{PPh}_3)_4(\text{SC}_{12})_4$ , and **b** fully-thiolated clusters. The insets show icosahedral (a) and truncated octahedral (b) units. The arrow indicates the fingerprint of the 2NN path in the closed packed structure (shown green in the inset) missing in the icosahedral clusters (a) but present in the truncated octahedral ones (b)



#### Nanoscale Atomic Clusters, Complexity of, Figure 6

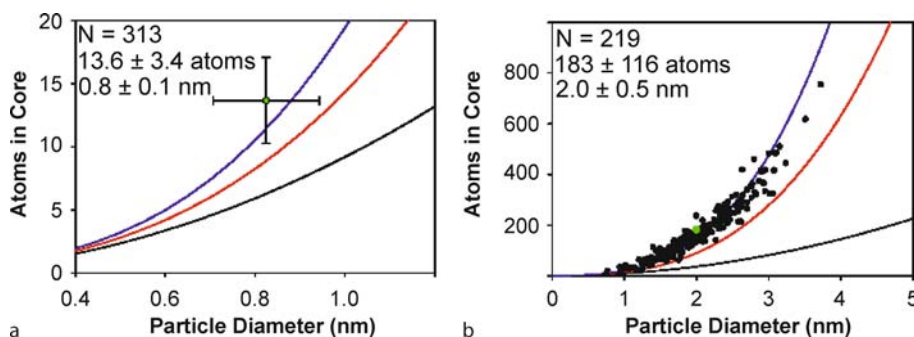
**a** High-resolution electron micrograph of the mixed-ligand clusters. The inset shows a filtered image of a single cluster with icosahedral structure. **b** Thiol-protected MPCs with cubic packing. The inset shows an indexed  $\mu$ -diffraction image taken from a single particle with a 2 nm diameter. Reprinted with permission from [59]

new mixed-ligand  $\text{Au}_{13}[\text{PPh}_3]_4[\text{SC}_{12}]_4$  as well as a monolayer protected cluster (MPC) system comprised of larger thiol-protected particles. The differences seen in the EXAFS data (Fig. 5) are profound and can be directly related to features of the atomic-level bonding present in each case. The mixed-ligand clusters adopt quasi-spherical shapes for their 13 atom cores (Figs. 6, 7). The full analyses unambiguously establish that an icosahedral geometry is present in each case, a bonding motif that stands in marked contrast with the *fcc* truncated octahedral structures adopted by the larger, fully thiolated MPCs (Fig. 5). These interpretations were independently verified by HRTEM and electron microdiffraction measurements (Fig. 6), and Z-contrast technique (Fig. 7) which identi-

fied the same structural motifs, albeit with less accurate detail [59,100].

While the combination of the experimental results points towards an icosahedral shape of the  $\text{Au}_{13}$  core, theoretical verification and a detailed interpretation of such a model was lacking. In particular, two central questions – ligand placement and anomalously high Au-Au bond length disorder – were left unanswered by the experimental results. First, EXAFS is not capable of discriminating between Au-S and Au-P neighbors, treating them cumulatively as Au-L ( $L = \text{S/P}$ ) pairs and obtaining the overall Au-L coordination number as the total number of Au-S and Au-P bonds divided by the total number of Au atoms. However, the phosphines and the thio-





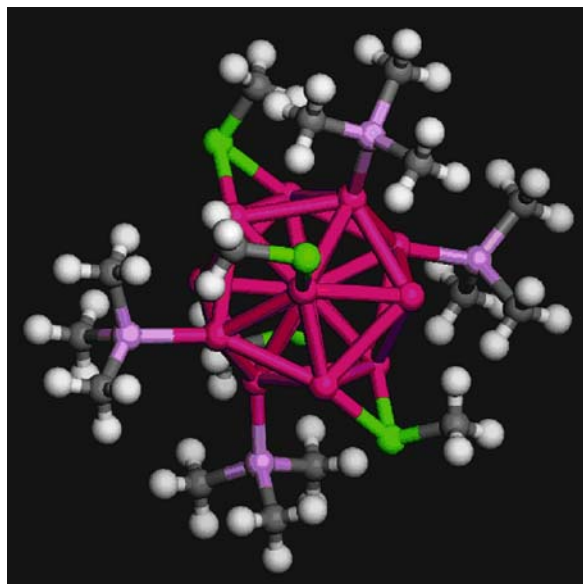
**Nanoscale Atomic Clusters, Complexity of, Figure 7**

Core atom counts for gold clusters measured using the quantitative HAADF-STEM technique for the mixed ligand (a) and thiolate-protected clusters (b), respectively, plotted against measured particle diameters. The blue, red and black lines are the spherical, hemispherical and monolayer island (111) models. The black and green dots are the data and the sample average. Reprinted with permission from [100]

lates may have distinctly different bonding motifs: on-top for phosphines [103] and bridge sites for thiolates [104]. Thus, the preferred ligand placement remains ambiguous. Second, the experimental distribution of Au-Au bond lengths was  $\sigma^2 = 0.017 \pm 0.005 \text{ \AA}^2$ , which is much larger than in bulk gold ( $0.008 \text{ \AA}^2$ ) at the same temperature (300 K) [59]. Such enhanced  $\sigma^2$  must be configurational in nature, because the temperature-dependent, dynamic component in nanoparticles has previously exhibited only weak, if any, size dependence [68]. However, neither EXAFS nor TEM provides enough information to uncover its origin.

Guliamov et al. [105] used the EXAFS and TEM results as a starting point for a theoretical analysis based on density functional theory (DFT) [106] of the mixed-ligand  $\text{Au}_{13}$  nanocluster,  $\text{Au}_{13}[\text{PPh}_3]_4[\text{S}(\text{CH}_2)_{11}\text{CH}_3]_4$ . For bare  $\text{Au}_{13}$  nanoclusters, many structures—including icosahedral [107], cuboctahedral [108], biplanar [109], and amorphous [110] ones—have been predicted theoretically as comprising the lowest energy configuration. Both ordered and disordered structures were predicted theoretically for ligated  $\text{Au}_{38}$  structures [111,112,113] and experimental verification was limited to scattering methods (e.g., x-ray diffraction) which are less than ideal for clusters of this size. This may reflect the existence of several energetically close isomers [113].

Calculations of different ligand structures relaxed from an initial icosahedral Au core geometry established that the symmetric bonded-ligand configuration, shown in Fig. 8, is energetically preferable. Here, the results of the calculations were found to be in good quantitative agreement with the EXAFS data: The mixed on-top and bridge thiol geometry was maintained, the average Au-Au bond length was  $2.88 \text{ \AA}$  with a standard deviation of  $\sigma^2 = 0.018 \text{ \AA}^2$ , the coordination number,  $N_{\text{Au-L}}$ , was 0.77, and



**Nanoscale Atomic Clusters, Complexity of, Figure 8**

Optimized structure of the icosahedral  $\text{Au}_{13}$  cluster. Red: Au; Green: S; Purple: P; Gray: C; White: H

the average Au-ligand distance was  $2.35 \text{ \AA}$ , i.e., in good agreement with all experimental results [59].

Further analysis of the ligand-induced deviation from the ideal icosahedral symmetry of the Au core reveals that the variation in the radial distances from the central atom to the other Au atoms is very small:  $R(\text{Au-Au})_{\text{rad}} = 2.78 \text{ \AA}$  with  $\sigma^2 = 0.005 \text{ \AA}^2$ . However, the in-shell tangential Au-Au bond lengths exhibit a much greater dispersion:  $R(\text{Au-Au})_{\text{tan}} = 2.92 \text{ \AA}$  with  $\sigma^2 = 0.017 \text{ \AA}^2$ , with the smallest distance found between two thiol-ligated Au atoms and the largest between one thiol-ligated and one free Au atom. This is consistent with the



strong covalent interaction expected between the S and Au atoms [114,115]. For comparison, we studied the radial versus tangential bond length distribution in the relaxed, bare icosahedral Au<sub>13</sub>. There, we found  $R(\text{Au-Au})_{\text{rad}} = 2.73 \text{ \AA}$  and  $R(\text{Au-Au})_{\text{tan}} = 2.87 \text{ \AA}$ .

The above results clearly reveal that the tangential strain induced by the ligands is much larger than the radial one. Both radial and tangential bond lengths in the ligated cluster are larger by  $\sim 1.8\%$  with respect to the bare one, but the induced dispersion in tangential bond lengths is much larger than in the radial ones. These findings can be interpreted via a combination of the asymmetry of the effective pair potential and the non-close-packed structure of the icosahedron. The ligands do not disorder the relatively stiff radial bonds, but do disorder atoms within the shell. The results suggest that it is the weakness of the in-shell Au-Au bonds, and thus lower energy penalty compared to the strong radial Au-Au bonds, that relieves otherwise strong stresses due to the asymmetry in the nature and bonding sites of thiolate and phosphine ligands.

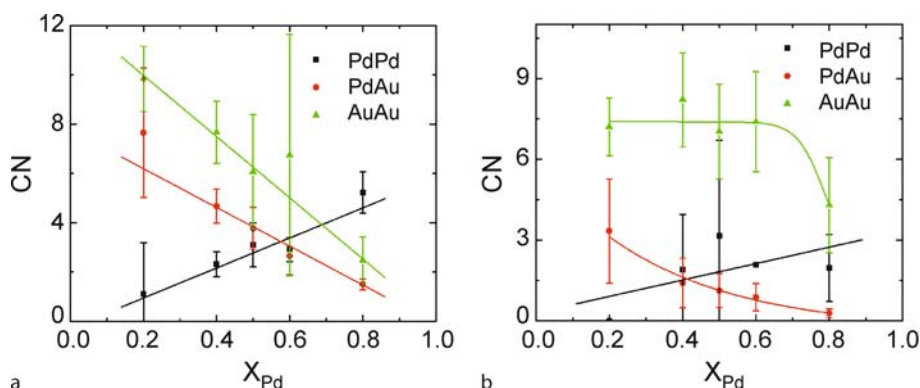
Experimental characterization of randomness of atomic distribution may be found by EXAFS. It is done by comparing experimental values of metal-metal coordination numbers against the model of (Fig. 3). For example, dendrimer-stabilized Pd/Au nanoalloys [116] are shown to be quasi-random or core-shell like, depending on the details of their preparation (Fig. 9).

This method is a powerful tool for quantifying short-range order in monodisperse clusters. However, a broad range of composition is required, to make such determination. If only one composition is available, the answer may be obtained from the combination of EXAFS, advanced electron microscopies (EDS, HREM, STEM) and DFT calculations. For example, EDS method revealed that the distribution of Pt:Ru was uniform, where each nanocluster

contained both Pt and Ru. Using the Z-contrast method for determining the number of atoms, Yang et al. [117] discovered that the average measured scattering cross-section corresponded to four PtRu<sub>5</sub> groups or 24 atoms. The diameters of these clusters were also measured from the STEM (Fig. 10) as well as by HREM, and the average was 15.6 Å. Using an averaged number density of Ru and Pt, if the shape of the cluster is spherical, then, for a diameter of 15 Å, the particle contains 21 PtRu<sub>5</sub> groups, whereas for a hemispherical shape, the cluster contains 11 PtRu<sub>5</sub> groups. These data shows that the particles are “raft-like” on the carbon black.

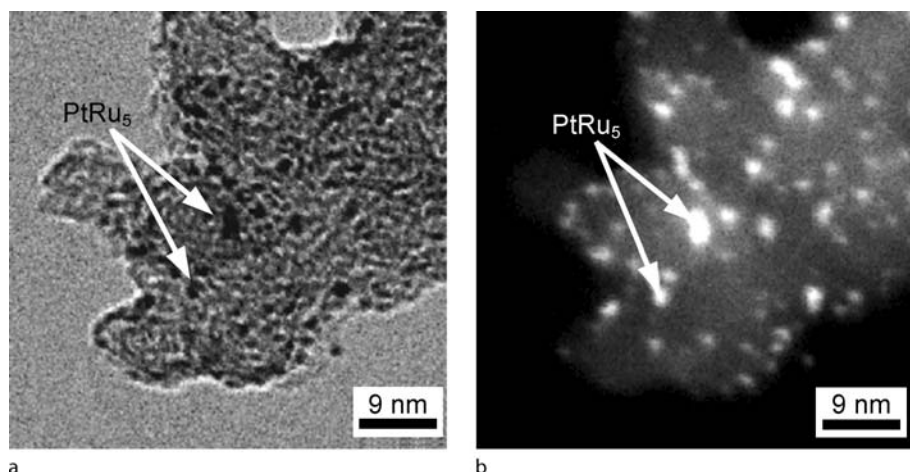
To demonstrate this point, Fig. 11 is a plot of the cluster diameter vs. the number of atoms per cluster for different 3-dimensional shapes. The two theoretical fits show the number of atoms for a (a) spherical shape with diameter,  $d$ , and (b) a hemispherical shape. Clearly, since the number of atoms is considerably less than would be predicted for a hemisphere, this demonstrates that the structure of the PtRu<sub>5</sub> is oblate on the carbon black support [117]. This result confirms the *truncated* cuboctahedral model proposed by Nashner et al. [94] based on results of their EXAFS analysis.

Wang et al. [118] confirmed EXAFS and electron microscopy results using DFT calculations, including the experimentally observed enhanced Pt-Pt bond length disorder. They revealed the origin of this disorder as due to the cluster/carbon-support interactions when samples are annealed in helium, whereas samples treated in hydrogen have disorder controlled by intraparticle effects, as discussed later. They explained the EXAFS observations that supported [PtRu<sub>5</sub>] metal clusters have fcc (111) cuboctahedral geometry and bulk-like metal-metal bond distances, even for small nanoparticles for which the average coordination number is much smaller than that in the



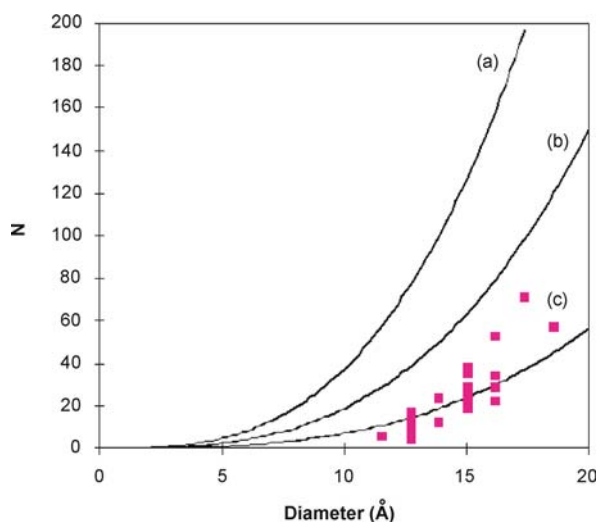
Nanoscale Atomic Clusters, Complexity of, Figure 9

Coordination numbers in nanoalloys: quasi-random dendrimer-stabilized Pd/Au alloys (a), and core-shell alloys (b). Samples courtesy of Crooks RM, U. of Texas at Austin



Nanoscale Atomic Clusters, Complexity of, Figure 10

BF image of PtRu<sub>5</sub> (a) and the corresponding HAADF image (b). Reprinted from [117]



Nanoscale Atomic Clusters, Complexity of, Figure 11

Plot of diameter vs. number of atoms for different shapes: a Sphere, b Hemisphere and c are the experimental data with the best fit where the 3-D aspect ratio is kept constant. Reprinted from [117]

bulk, and that Pt in the bimetallic clusters segregates to the top (111)-layer of the nanoparticle, as hypothesized based on EXAFS results.

### Structural Relaxation of Nanoparticles

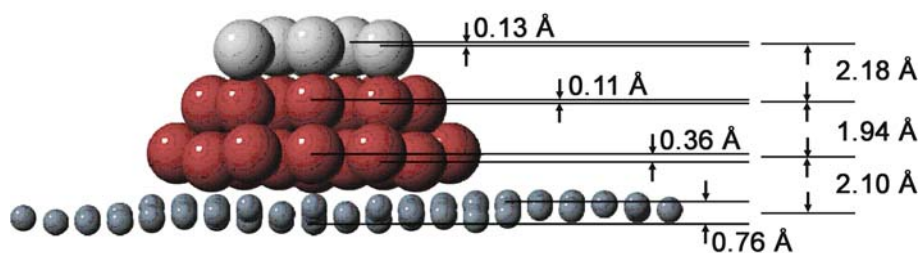
The data discussed above illustrate one level of new knowledge that comes from direct measurements of atomic scale structure. It is useful, though, to look at additional levels of insight that can be developed from data of this sort. Metal nanoclusters contain many distinct types of atoms – the

various habits found at the cluster surfaces, those residing at the metal support interface or cluster core, etc. [119] The metal-metal bonding, as a result of this diversity, is present as an ensemble embedding substantial configurational, static (as opposed to vibrational, dynamic) disorder.

### Structural Relaxation in Freestanding and Supported Clusters

The case of the icosahedral Au clusters provides an interesting example in this regard. In this case, one moment of this effective disorder in the first shell bonding can be isolated and this in turn enables one to evaluate separately the nature of the icosahedral strain that it embeds. The latter parameter is one that directly illustrates the nature of the energetics that selects such non-bulk habits for the smallest clusters. The EXAFS analysis (Fig. 5a, inset) revealed the presence of an icosahedral strain of ca. 2.5% as compared to the 5% strain predicted by geometrically. This strain relaxation is correlated with and can be understood in terms of the ligand bonding – the simplest model of a support interaction (see below) – that terminate the cluster's surface. Obtaining such levels of structural understanding, however, is only possible if the synthetic protocols used deliver samples that are comprised of extremely monodispersed nanoparticles, a condition where it becomes possible to discriminate intra vs inter particle disorder (where the latter arises as a result of size-dependent atomic bond-length relaxation [61,85]).

It is also now well understood that more general forms of structural relaxations are common to the energy landscapes of supported metal clusters. The smallest clusters,

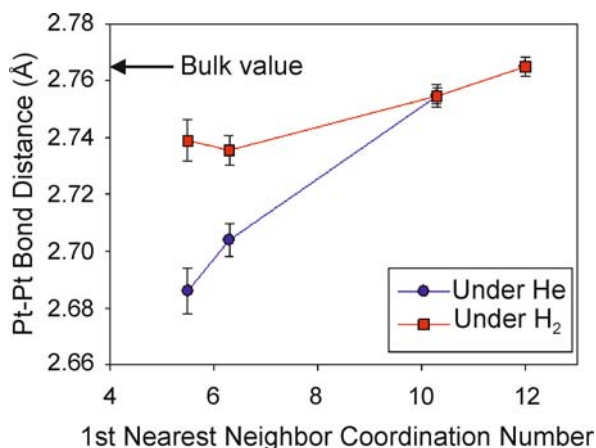


#### Nanoscale Atomic Clusters, Complexity of, Figure 12

Side view of the relaxed structures calculated with DFT-LDA for PtRu clusters supported on a graphite surface. The *small dark sphere* stands for C atoms, *large light (dark) spheres* for Pt (Ru) atoms. Both the intra-layer buckling and inter-layer distance are listed. Reprinted with permission from [118]

for example, are known to show contractions of their metal-metal bonding distances, [85] an outcome that can be understood intuitively in the context of the large number of non-bulk like atoms that would be found in a cluster that is 1 nm in size – the relative fractions of surface and support-bonded atoms being large in such cases.

The data in Fig. 13 illustrate this point in greater detail, highlighting EXAFS measurements made on a series of rigorously characterized Pt catalysts supported on  $\gamma$ - $\text{Al}_2\text{O}_3$  [9]. As synthesized, the cluster distributions of atomic mass were held to exceptionally low values (a property stiffly characterized by microscopy). The data show the strong size dependent scaling of the average first-shell Pt-Pt bonding distances found in these samples. These distances are plotted here in terms of the number of first shell neighbors found experimentally by EXAFS, a parameter that explicitly correlates with both the particle's shape and diameter (as discussed for the data presented in Fig. 3, and for these samples showing by microscopy diameters

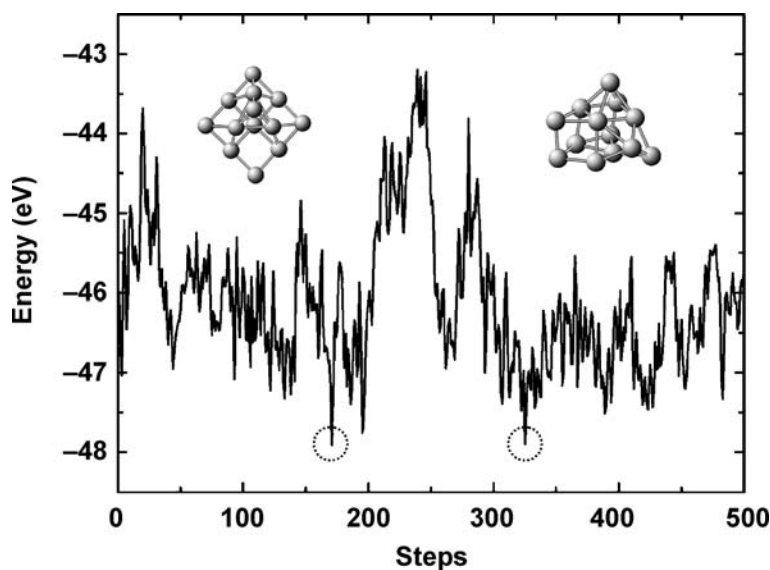


#### Nanoscale Atomic Clusters, Complexity of, Figure 13

The size-correlated average first shell M-M bonding distances for a series of Pt catalysts supported on  $\gamma$ - $\text{Al}_2\text{O}_3$

spanning a range from 0.9 to 2.8 nm). The sample with the smallest Pt particles (0.9 nm) consisted of a highly homogeneous sample of  $\text{Pt}_{15}$  clusters (HAADF-STEM). The EXAFS data illustrate that size effects, in this case, elicit strong contractions of the metal-metal bonding distances – a structural relaxation that also develops a correlated and significant degree of interatomic disorder in the catalytic clusters. These relaxations are responsive to the presence of adsorbates – the data in the figure illustrate that hydrogen, which dissociates and passivates the surface bonds of the clusters lifts the relaxations (albeit not fully to bulk values for all but the largest clusters). The larger body of data implicitly suggests a model for the structural relaxations in which the bonding present at the ambient cluster surfaces/interfaces play an exceptionally important role – establishing the biases seen in the average distances and effective disorder measured by EXAFS.

Advances made in the first principle theory and MD simulations allow one to investigation of structural relaxation and relative stability of various cluster morphologies. Nonetheless, DFT-based simulations must be carefully validated before use in each system. For example, Wang and Johnson [120] have shown that  $N$ -atom (for  $N \leq 20$ ) cluster morphologies are highly sensitive to the DFT exchange-correlation functional, and standard generalized gradient corrected, local (spin) density approximation often yield incorrect structural properties and energetics compared to high-level quantum chemistry. In addition, with the numerous geometrical and chemical configurations associated with finding ground state and lowest-energy excited-state morphologies (e. g., as in Fig. 1), one cannot be assured with absolute certainty that a simulation has found the absolute ground state for comparison to observation because the simulation times are short (order of nanoseconds to pico-seconds) and starting geometries are few. In this regard, the most commonly employed global-optimization strategies are simulated annealing (sometimes combined, e. g., see [4] with empirical



**Nanoscale Atomic Clusters, Complexity of, Figure 14**

First-principles MD global (potential energy surface) optimization for 500 time steps for a  $\text{Pt}_{13}$  cluster that starts with an  $\text{O}_h$  cluster. Simulation was held at 2000 K for 4 to 10 picoseconds at time steps of 20 femtoseconds and used a single k-point in a periodic box with large vacuum. Two low-energy basins are circled, with their corresponding structures shown in insets after full relaxation of the cluster. All structures in Fig. 1 were found using this approach. Reprinted figure with permission from [3]

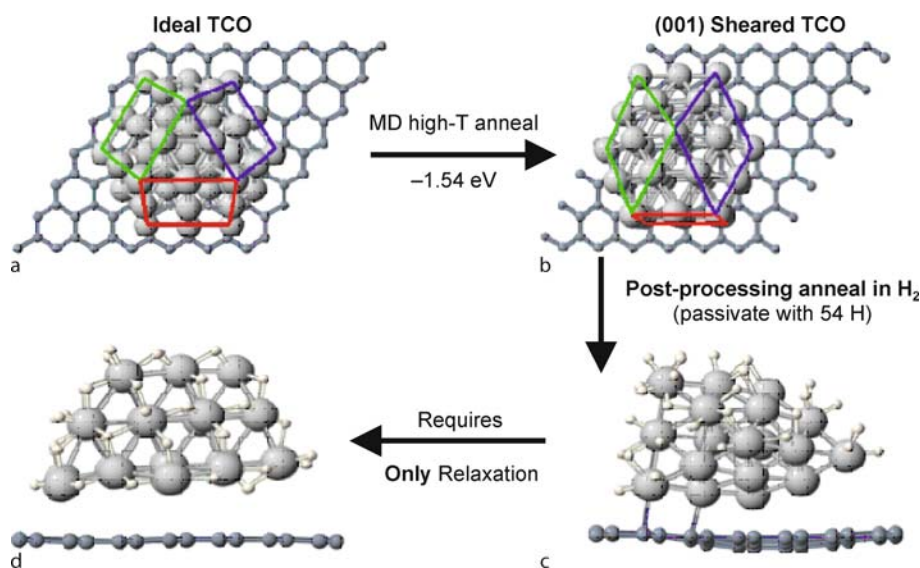
force-fields not containing proper cluster correlation effects), genetic and evolutionary algorithms (e. g., see example review [121] and with improved GAs [122]), and variants of basin-hopping that utilize temperature as a tunable parameter to sample large, but not all, portions of solution space (e. g., Monte Carlo [123] and molecular dynamics [3,4] using very small (not converged) k-meshes). Genetic algorithms and basin hopping are comparable in efficiency, both superior to simulated annealing. An example of basin hopping within MD is shown in Fig. 14, and used to obtain results in Fig. 1.

An illustrative example of DFT studies of structural relaxation and morphological transitions in nanoparticles is shown in Fig. 15, where Wang et al. [118] were able to characterize fully the adsorbate-dependent morphological habits of a model  $\text{Pt}_{37}$  cluster supported on carbon. Importantly, for comparison to experiment, consideration was made for the annealing stages in the synthesis, including post-processing in hydrogen atmosphere before characterization. With an anneal in hydrogen, the preferred structural habit for the cluster is a truncated cuboctahedron with large Pt-Pt bond length disorder in the bond distribution (from 2.62 to 2.93 Å) that arises from structural anisotropy within a cluster. This disorder is highest for bare clusters and is reduced when hydrogen atoms are allowed to adsorb. The disorder seen in each case provides a near quantitative agreement with values deduced from experiment [118,124]. These DFT calculations further re-

veal that the  $\text{Pt}_{37}$  truncated cuboctahedron (TCO, shown in Fig. 15a) has a morphological instability driven by the shearing of (100) to (111) facets to lower the surface energy (see Fig. 15b) – a remnant of the electronic mechanism responsible for (100)-surface reconstruction in semi-infinite bulk Pt. [124] However, with H passivation, this shear instability is removed and the TCO is highly stable, [124] as observed in experiments [94] and which is clearly a result of the annealing in  $\text{H}_2$ .

Theory also predicts that M-M bond length contractions occur – ones that vary substantially with the presence of adsorbate bonding. They found, for example, that the DFT derived average first nearest neighbor Pt-Pt bond length for the model  $\text{Pt}_{37}$  cluster increases by 3% with H passivation (2.68 Å without H to 2.75 Å with H) [124]. These values provide a remarkable agreement with those determined experimentally (EXAFS 2.75 Å) when samples are annealed in hydrogen [68] versus helium [9]. Importantly, H-passivation effect on the Pt-Pt bond lengths are strongly dependent on the size of the cluster (both in simulations and observed); for smaller Pt nanoparticles with 1st shell coordination number (CN) from 5 to 7, an increase in Pt-Pt bond lengths is observed on various supports, whereas, for larger clusters with CN beyond 8 (i. e., approaching bulk CN of 12), no such effect is observed. The reason being that the relation between 1st shell CN and particle size is not linear, but roughly proportional to the  $\sqrt{D}$ , the particle diameter ( $D = 2R$ ).





### Nanoscale Atomic Clusters, Complexity of, Figure 15

Synopsis of the calculated DFT morphological changes of the structure of a Pt<sub>37</sub> cluster on a carbon support during high-temperature anneal and post-processing anneal in H<sub>2</sub> performed during synthesis of real samples, see [124] for details. **a** High-T anneal of the ideal cuboctahedral (TCO) cluster, which **b** lowers its energy by internal shear of all (100) facets (shown by the red, green and blue planes) such that they resemble (111) facets, a mechanism which is a remnant of the electronically-driven shear instability for surface reconstruction in semi-infinite bulk Pt. For direct comparison to experiment, **c** the structure found in simulated post-processing anneal in H<sub>2</sub>, which **d** structurally reverts simple by relaxation (downhill in energy) to the observed cuboctahedral structure, showing then negligible interaction with the carbon support. For structural comparison to experiment, see text. Reprinted with permission from [118]

Clearly, careful implementation and validation of theory not only shows excellent agreement with a range physical properties (from bond-length distributions, structural parameters, electronic properties, etc.) but also reveals the key physics controlling these manifested behaviors, including the important of cluster-support interactions and charge transfer that impacts the structural and concomitant catalytic behavior, so highly prized for technological purposes.

### Future Directions

With respect to the atomic arrangements present in nanoparticles, the availability of a new state of the art instruments for studies by electron microscopy and x-ray absorption spectroscopy provides an exceptional opportunity to gain new levels of understandings of the structural motives present in these systems. We believe that direct imaging of the 3D atomic arrangement present in such clusters will become eventually possible. When the future synchrotron sources with exceptional brightness (equivalent to the counting statistics of 10<sup>12</sup> photons/sec at the 1 nm focal spot) become available, it will become possible to measure structural dynamics in individual nanopar-

ticles by EXAFS. With the newest electron microscopes (e.g., JEOL (JEM) 2200F), high-resolution electron microscopy imaging will be possible with dramatically better resolution and, hence, improved interpretability. Since the spherical aberration causes delocalization in the images proportional to  $\Theta^3$ , where  $\Theta$  is the Bragg angle of the diffracted beam, the reduction of the spherical aberration will eliminate the presence of “ghost” images and/or fringes, leading to significantly improved clarity of images. The focal series reconstruction of exit wavefunction can further improve the HREM interpretability. Because of the ability to obtain spectral information atomic-by-atomic column, changes in the EELS signal across a single nanoparticle, from the surface to the interface, will be achievable. This will provide detailed information about individual arrangement of atoms within an individual particle, such as core-shell structure or surface segregation. Furthermore, with the significantly improved HAADF-STEM imaging resolution to 1 Å, significantly better than the 1 nm<sup>3</sup> tomography resolution available on modern non-aberration corrected instruments [125]. Additional possibilities that are being engendered by methods of coherent electron diffraction further suggest that it may be possible to vastly exceed the capabilities afforded by even



the very advanced capabilities of the JEOL 2200F instrument for electron tomography. For example, via over-sampling and phase retrieval it will be possible to invert a measured single cluster nano-diffraction pattern to obtain an atomic resolution 3D structure directly [126].

Current theoretical methods will be extended by enabling DFT-based calculations to identify low-energy structures, utilizing a rapid – optimization and selection protocols that appear to be well suited to applications involving supported nanoscale materials. An example is the DFT-based MD approach, in which the simulated nanocluster is heated to high temperature (e.g., 2000 K) and simulated rapidly using a coarse k-space mesh, is used to identify (relatively) rapidly several low-energy conformations, see Figs. 1 and 14. These conformations are then revisited with DFT-MD with a fine k-space mesh and configurations are then identified after full ionic relaxations. In the future, such protocols will be integrated with improved DFT exchange-correlation functionals to explore the nature of the energy landscapes that define the accessible and correct structures for a large variety of classes of nanoscale materials. Note that improved DFT exchange-correlation functionals [127,128,129,130,131,132,133,134,135] (such as hybrid B3LYP or PBE0, WC, and HSC) are computationally much more expensive than standard gradient corrected or local density approximation but yield much more correct results for nanoparticles, surface interactions, and some solid phases, such as ferroelectronics, which are highly sensitive to volume errors.

Complexity of nanoscale clusters can be quantitatively understood. However, due to the complexity of the characterization approaches that such understanding demands, only self-consistent, integrated, multi-disciplinary analysis methodologies appear to be heading in the direction of systematically solving their mysteries.

## Acknowledgments

We are grateful to S. Sanchez, L. Li, A. Urban, L.-L. Wang, F. Dukesz and Q. Wang for help in preparing this manuscript. We would like to thank R.M. Crooks, M. Knecht and M. Meir for sharing samples used in these studies. We acknowledge support by the US Department of Energy (DOE) Grant No. DE-FG02-03ER15476.

## Bibliography

### Primary Literature

- Lai SL, Guo JY, Petrova V, Ramanath G, Allen LH (1997) Size-dependent melting properties of small tin particles: nanocalorimetric measurements. *Phys Rev Lett* 77:99–102
- Jesser WA, Shneck RZ, Gile WW (2004) Solid-liquid equilibria in nanoparticles of Pb-Bi alloys. *Phys Rev B* 69(14):144121
- Wang LL, Johnson DD (2007) Density functional study of structural trends for late-transition-metal 13-atom clusters. *Phys Rev B* 75:235405
- Futschek T, Marsman M, Hafner J (2005) Structural and magnetic isomers of small Pd and Rh clusters: an ab initio density functional study. *J Phys Condens Matter* 17:5927–5963
- Futschek T, Hafner J, Marsman M (2006) Stable structural and magnetic isomers of small transition-metal clusters from the Ni group: an ab initio density-functional study. *J Phys Condens Matter* 18:5703–5748
- Ferrando R, Jellinek J, Johnston RL (2008) Nanoalloys: From theory to applications of alloy clusters and nanoparticles. *Chem Rev* 108(3):845–910
- Baletto F, Ferrando R (2005) Structural properties of nanoclusters: Energetic, thermodynamic, and kinetic effects. *Rev Mod Phys* 77:371–423
- Daniel MC, Astrue D (2004) Gold nanoparticles: assembly, supramolecular chemistry, quantum-size-related properties, and applications toward biology, catalysis, and nanotechnology. *Chem Rev* 104(1):293–346
- Kang JH, Menard LD, Nuzzo RG, Frenkel AI (2006) Unusual non-bulk properties in nanoscale materials: thermal metal-metal bond contraction of  $\gamma$ -alumina-supported Pt catalysts. *J Am Chem Soc* 128:12068–12069
- Kamat PV (1993) Photochemistry on nonreactive and reactive (semiconductor) surfaces. *Chem Rev* 93:267–300
- Gates BC (1995) Supported metal clusters: synthesis, structure, and catalysis. *Chem Rev* 95:511–522
- Shahbazyan TV, Perakis IE (1999) Size-dependent correlation effects in the ultrafast optical dynamics of metal nanoparticles. *Phys Rev B* 60(12):9090–9099
- Tsunoyama H, Sakurai H, Tsukuda T (2006) Size effect on catalysis of gold clusters dispersed in water for aerobic oxidation of alcohol. *Chem Phys Lett* 429:528–532
- Roucoux A, Schults J, Patin H (2002) Reduced transition metal colloids: A novel family of reusable catalysts. *Chem Rev* 102:3757–3778
- Schmid G (2006) Metal Nanoparticles, Synthesis of. In: *Encyclopedia of Inorganic Chemistry*, 2nd edn, vol 5. Wiley-VCH, Weinheim
- Taneja R, Chandra R, Banerjee R, Ayyub P (2001) Structure and properties of nanocrystalline Ag and Cu<sub>2</sub>O synthesized by high pressure sputtering. *Scripta Mater* 44:1915–1918
- Spadavecchia J, Prete P, Lovergine N, Tapfer L, Rella R (2005) Au nanoparticles prepared by physical method on Si and sapphire substrates for biosensor applications. *J Phys Chem B* 109:17347–17349
- Mafune F, Jun-ya K, Takeda Y, Kondow T (2002) Full physical preparation of size-selected gold nanoparticles in solution: laser ablation and laser-induced size control. *J Phys Chem B* 106(31):7575–7577
- Ramachandra Rao CN, Kulkarni GU, Thomas PJ, Peter P Edwards (2000) Metal nanoparticles and their assemblies. *Chem Soc Rev* 29:27–35
- Giesen B, Orthner HR, Kowalik A, Roth P (2004) On the interaction of coagulation and coalescence during gas-phase synthesis of Fe-agglomerates. *Chem Eng Sci* 59:2201–2211

21. Kamat PV, Flumiani M, Hartland GW (1998) Picosecond dynamics of silver nanoclusters. Photoejection of electrons and fragmentation. *J Phys Chem B* 102:3123–3128
22. Kurita H, Takami A, Koda S (1998) Size reduction of gold particles in aqueous solution by pulsed laser irradiation. *Appl Phys Lett* 72:789–791
23. Takami A, Kurita H, Koda S (1999) Laser-induced size reduction of noble metal particles. *J Phys Chem B* 103(8):1226–1232
24. Hirai H, Nakao Y, Toshima N (1978) Colloidal rhodium in poly(vinylpyrrolidone) as hydrogenation catalyst for internal olefins. *Chem Lett* 7(5):545–548
25. Kiwi J, Gratzel M (1979) Protection size factors, and reaction dynamics of colloidal redox catalysis mediating light induced hydrogen evolution from water. *J Am Chem Soc* 101:7214–7127
26. Chechik V, Crooks RM (2000) Dendrimer-encapsulated Pd nanoparticles as fluorous phase-soluble catalysts. *J Am Chem Soc* 122(6):1243–1244
27. Harriman A, Thomas JM, Millward GR (1987) Catalytic and structural properties of iridium-iridium dioxide colloids. *New J Chem* 11:757–762
28. Zhu J, Shen Y, Xie A, Qiu L, Zhang Q, Zhang S (2007) Photoinduced synthesis of anisotropic gold nanoparticles in room-temperature ionic liquid. *J Phys Chem C* 111(21):7629–7633
29. McGilvray KL, Decan MR, Wang D, Scaiano JC (2006) Facile photochemical synthesis of unprotected aqueous gold nanoparticles. *J Am Chem Soc* 128:15980–15981
30. Yang S, Wang Y, Qingfeng W, Zhang R, Ding B (2007) Colloids *Surf A: Physicochem Eng Aspects* 301:174–183
31. Mallick K, Witcomb MJ, Scurrall MS (2004) Polymer stabilized silver nanoparticles: A photochemical synthesis route. *J Mat Sci* 39(14):4459–4463
32. Lee J, Ryu J, Choi W (2007) Synthesis of gold and platinum nanoparticles using visible light activated Fe(III)-complex. *Chem Lett* 36:176–177
33. Pei L, Mori K, Adachi M (2004) Formation process of two-dimensional networked gold nanowires by citrate reduction of  $\text{AuCl}_4^-$  and the shape stabilization. *Langmuir* 20(18):7837–7843
34. Suslick KS, Choe SB, Cichowlas AA, Grinstaff MW (1991) Sonochemical synthesis of amorphous iron. *Nature* 353:414–416
35. Didenko YT, McNamara WB, Suslick KS (1999) Hot spot conditions during cavitation in water. *J Am Chem Soc* 121:5817–5818
36. Su CH, Wu PL, Yeh CS (2003) Sonochemical synthesis of well-dispersed gold nanoparticles at iced temperature. *J Phys Chem B* 107:14240–14243
37. Jiang LP, Wang AN, Zhano Y, Zhang JR, Zhu JJ (2004) A novel route for the preparation of monodisperse silver nanoparticles via a pulsed sonoelectrochemical technique. *Inorg Chem Comm* 7:506–509
38. Dhas NA, Gedanken A (1998) Sonochemical preparation and properties of nanostructured palladium metallic clusters. *J Mater Chem* 8(2):445–450
39. Kan C, Cai W, Li C, Zhang L, Hofmeister H (2003) Ultrasonic synthesis and optical properties of Au/Pd bimetallic nanoparticles in ethylene glycol. *J Phys D* 36:1609–1614
40. Fujimoto T, Terauchi SY, Umehara H, Kojima I, Henderson W (2001) Sonochemical preparation of single-dispersion metal nanoparticles from metal salts. *Chem Mater* 13:1057–1060
41. Huang S, Ma H, Zhang X, Yong F, Feng X, Pan W, Wang X, Wang Y, Chen S (2005) Electrochemical synthesis of gold nanocrystals and their 1D and 2D organization. *J Phys Chem B* 109(42):19823–19830
42. Rodriguez-Sanchez L, Blanco MC, Lopez-Quintela MA (2000) Electrochemical synthesis of silver nanoparticles. *J Phys Chem B* 104:9683–9688
43. Starowicz M, Stypula B, Banas J (2006) Electrochemical synthesis of silver nanoparticles. *Electrochem Comm* 8(2):227–230
44. Zhou M, Chen S, Ren H, Wu L, Zhano S (2005) Electrochemical formation of platinum nanoparticles by a novel rotating cathode method. *Physica E* 27(3):341–351
45. Ueda M, Dietz H, Anders A, Knepe H, Meixner A, Plieth W (2002) Double-pulse technique as an electrochemical tool for controlling the preparation of metallic nanoparticles. *Electrochem Acta* 48:377–386
46. Plieth W, Dietz H, Anders A, Sandmann G, Meixner A, Weber M, Knepe H (2005) Electrochemical preparation of silver and gold nanoparticles: Characterization by confocal and surface enhanced Raman microscopy. *Surf Sci* 597:119–126
47. Toshima N, Yonezawa T, Kushihashi K (1993) Polymer-protected palladium–platinum bimetallic clusters: preparation, catalytic properties and structural considerations. *J Chem Soc* 89:2537–2543
48. Toshima N, Harada M, Yamazaki Y, Asakura K (1992) Catalytic activity and structural analysis of polymer-protected gold-palladium bimetallic clusters prepared by the simultaneous reduction of hydrogen tetrachloroaurate and palladium dichloride. *J Phys Chem* 96:9927–9933
49. Han S W, Kim Y, Kim K (1998) Dodecanethiol-derivatized Au/Ag bimetallic nanoparticles: TEM, UV/VIS, XPS, and FTIR analysis. *J Colloid Interface Sci* 208:272–278
50. Link S, Wang ZL, El-Sayed MA (1999) Alloy formation of gold-silver nanoparticles and the dependence of the plasmon absorption on their composition. *J Phys Chem B* 103:3529–3533
51. Sun S, Murray CB, Weller D, Folks L, Moser A (2000) Monodisperse FePt nanoparticles and ferromagnetic FePt nanocrystal superlattices. *Science* 287:1989–1992
52. Bian B, Hirotsu Y, Sato K, Ohkubo T, Makino A (1999) Structures and magnetic properties of oriented Fe/Au and Fe/Pt nanoparticles on  $\alpha\text{-Al}_2\text{O}_3$ . *J Electron Microsc* 48:753–758
53. Liu HB, Pal U, Medina A, Maldonado C, Ascencio JA (2005) Structural incoherency and structure reversal in bimetallic Au-Pd nanoclusters. *Phys Rev B* 71:075403
54. Dadge JW, Islam M, Dharmadhikari AK, Mahamuni SR, Aiyer RC (2006) Hyper-Rayleigh scattering in electrochemically synthesized Ag–Au coupled clusters. *J Phys Condens Matter* 18:5405–5413
55. Mandal M, Kundu S, Ghosh SK, Pal T (2004) Micelle-mediated UV-photoactivation route for the evolution of  $\text{Pd}_{\text{core}}\text{-Au}_{\text{shell}}$  and  $\text{Pd}_{\text{core}}\text{-Ag}_{\text{shell}}$  bimetallics from photogenerated Pd nanoparticles. *J Photochem Photobiol A* 167:17–22
56. Hiraoka K, Toshima N (2003) Ag/Rh bimetallic nanoparticles formed by self-assembly from Ag and Rh monometallic nanoparticles in solution. *Chem Lett* 32(1):78–79
57. Kanemaru M, Shiraishi Y, Koga Y, Toshima N (2005) Calorimetric study on self-assembling of two kinds of monometallic nanoparticles in solution. *J Therm Analysis Calorim* 81:523–527

58. Wang Y, Toshima N (1997) Preparation of Pd-Pt bimetallic colloids with controllable core/shell structures. *J Phys Chem B* 101:5301–5303
59. Menard LD, Xu H, Gao S, Twisten RD, Harper AS, Song Y, Wang G, Douglas AD, Yang JC, Frenkel AI, Murray RW, Nuzzo RG (2006) Metal core bonding motifs of monodisperse icosahedral Au<sub>13</sub> and larger Au monolayer-protected clusters as revealed by x-ray absorption spectroscopy and transmission electron microscopy. *J Phys Chem B* 110(30): 14564–14573
60. Wang X, Hanson JC, Frenkel AI, Kim JJ, Rodriguez JA (2004) Time-resolved studies for the mechanism of reduction of copper oxides with carbon monoxide: Complex behavior of lattice oxygen and the formation of suboxides. *J Phys Chem B* 108:13667
61. Frenkel AI, Nemzer S, Pister I, Soussan L, Harris T, Sun Y, Rafailovich MH (2005) Size-controlled synthesis and characterization of thiol-stabilized gold nanoparticles. *J Chem Phys* 123:184701–6
62. Sun Y, Frenkel AI, Isseroff R, Shonbrun C, Forman M, Shin K, Koga T, White H, Zhang L, Zhu Y, Rafailovich MH, Sokolov JC (2006) Characterization of palladium nanoparticles by using x-ray reflectivity, EXAFS, and electron microscopy. *Langmuir* 22(2):807–816
63. Frenkel AI, Frankel SC, Liu T (2005) Structural stability of giant polyoxomolybdate molecules as probed by EXAFS. *Phys Scripta* T115:721–723
64. Frenkel AI, Menard LD, Northrup P, Rodriguez JA, Zypman F, Glasner D, Gao SP, Xu H, Yang JC, Nuzzo RG (2007) Geometry and charge state of mixed-ligand Au<sub>13</sub> nanoclusters. *AIP Conf Proc* 882:749–751
65. See, e. g., Bonačić-Koutecký V, Fantucci P, Koutecký V (1991) Quantum chemistry of small clusters of elements of groups Ia, Ib, and IIa: fundamental concepts, predictions, and interpretation of experiments. *Chem Rev* 91(5):1035–1108
66. Pyykkö R (2004) Theoretical chemistry of gold. *Angew Chem Int Ed* 43:4412–4456
67. Gurman SJ (1995) Interpretation of EXAFS data. *J Synchrotron Rad* 2:56–63
68. Frenkel A, Hills C, Nuzzo R (2001) A view from the inside: Complexity in the atomic scale ordering of supported metal nanoparticles. *J Phys Chem B* 105(51):12689–12703 (Feature Article)
69. Zabinsky SI, Rehr JJ, Ankudinov A, Albers RC, Eller MJ (1995) Multiple-scattering calculations of x-ray-absorption spectra. *Phys Rev B* 52:2995–3009
70. Newville M (2001) IFEFFIT: interactive XAFS analysis and FEFF fitting. *J Synchrotron Rad* 8:322–324
71. Montejano-Carrizale JM, Aguilera-Granja F, Moran-Lopez JL (1997) Direct enumeration of the geometrical characteristics of clusters. *NanoStruct Mater* 8(3):269–287
72. Calvin S, Miller MM, Goswami R, Cheng SF, Mulvaney SP, Whitman LJ, Harris VG (2003) Determination of crystallite size in a magnetic nanocomposite using extended x-ray absorption fine structure. *J Appl Phys* 94(1):778–783
73. Lytle FW, Via GH, Sinfelt JH (1977) New application of extended x-ray absorption fine structure (EXAFS) as a surface probe-nature of oxygen interaction with a ruthenium catalyst. *J Chem Phys* 67:3831–3832
74. Sinfelt JH, Via GH, Lytle FW (1977) Extended x-ray absorption fine structure (EXAFS) of supported platinum catalysts. *J Chem Phys* 68:2009–2010
75. Via GH, Sinfelt JH, Lytle FW (1979) Extended x-ray absorption fine structure (EXAFS) of dispersed metal catalysts. *J Chem Phys* 71:690–699
76. Sinfelt JH, Via GH, Lytle FW (1980) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Ru–Cu clusters. *J Chem Phys* 72:4832–4844
77. Sinfelt JH, Via GH, Lytle FW, Greegor RB (1981) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Os–Cu clusters. *J Chem Phys* 75:5527–5537
78. Sinfelt JH, Via GH, Lytle FW (1982) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Pt–Ir clusters. *J Chem Phys* 76:2779–2789
79. Marques EC, Sandstrom DR, Lytle FW, Greegor RB (1982) Determination of thermal amplitude of surface atoms in a supported Pt catalyst by EXAFS spectroscopy. *J Chem Phys* 77(2):1027–1034
80. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Rh–Cu clusters. *J Chem Phys* 78(2):882–889
81. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Ir–Rh clusters. *J Chem Phys* 78(5): 2533–2541
82. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Re–Cu, Ir–Cu, and Pt–Cu clusters. *J Chem Phys* 83(1):353–360
83. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Ag–Cu and Au–Cu clusters. *J Chem Phys* 83(9):4793–4799
84. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) of Pt–Re and Pd–Re clusters. *J Chem Phys* 87(11):6354–6363
85. Mays CW, Vermaak JS, Kuhlmann-Wilsdorf D (1968) On surface stress and surface tension: II. Determination of the surface stress of gold. *Surf Sci* 12(2):134–140
86. Montano PA, Schulze W, Tesche B, Shenoy GK, Morrison TI (1984) Extended x-ray-absorption fine-structure study of Ag particles isolated in solid argon. *Phys Rev B* 30(2):672–677
87. Montano PA, Purdum H, Schenoy GK, Morrison TI, Schulze W (1985) X-ray absorption fine structure study of small metal clusters isolated in rare-gas solids. *Surf Sci* 156:228–233
88. Montano PA, Shenoy GK, Alp EE, Schulze W, Urban J (1986) Structure of copper microclusters isolated in solid argon. *Phys Rev Lett* 56(19):2076–2079
89. Jiang P, Jona F, Marcus PM (1987) Surface effects in metal microclusters. *Phys Rev B* 36(12):6336–6338
90. Finnis MW, Heine V (1974) Theory of lattice contraction at aluminium surfaces. *J Phys F* 4(4):L37–L41
91. Frenkel AI (1999) Solving the structure of nanoparticles by multiple-scattering EXAFS analysis. *J Synchrotron Rad* 6: 293–295
92. Hwang BJ, Sarma LS, Chen JM, Chen CH, Shih SC, Wang GR, Liu DG, Le JF, Tang M (2005) Structural models and atomic distribution of bimetallic nanoparticles as investigated by x-ray absorption spectroscopy. *J Am Chem Soc* 127(31): 11140–11145

93. Cowley JM (1965) Short-range order and long-range order parameters. *Phys Rev* 138(5A):A1384–A1389
94. Nashner MS, Frenkel AI, Adler DL, Shapley JR, Nuzzo RG (1997) Structural characterization of carbon-supported platinum-ruthenium nanoparticles from the molecular cluster precursor  $\text{PtRu}_5\text{C}(\text{Co})_{16}$ . *J Am Chem Soc* 119:7760–7771
95. Singhal A, Yang JC, Gibson JM (1997) STEM-based mass spectroscopy of supported Re clusters. *Ultramicroscopy* 67: 191–206
96. Thomas JM, Midgley PA (2004) High-resolution transmission electron microscopy. The ultimate nanoanalytical technique. *Chem Commun* 7:1253–1267
97. Ishizuka K (2004) FFT multislice method—the silver anniversary. *Microsc Microanal* 10(1):34–40
98. Liu J (2005) Scanning transmission electron microscopy and its application to the study of nanoparticles and nanoparticle systems. *J Electron Microsc* 54(3):251–278
99. Liu W (2007) Multi-scale catalyst design. *Chem Eng Sci* 62(13):3502–3512
100. Menard LD, Gao S, Xu H, Twisten RD, Harpe AS, Song Y, Wang G, Douglas AD, Yang JC, Frenke AI, Nuzzo RG, Murray RW (2006) Sub-nanometer Au monolayer-protected clusters exhibiting molecule-like electronic behavior: Quantitative high-angle annular dark-field scanning transmission electron microscopy and electrochemical characterization of clusters with precise atomic stoichiometry. *J Phys Chem B* 110(26):12874–12883
101. Patterson AL (1939) The Scherrer formula for x-ray particle size determination. *Phys Rev* 56(10):978–982
102. Menard LD, Xu F, Nuzzo RG, Yang JC (2006) Preparation of  $\text{TiO}_2$ -supported Au nanoparticle catalysts from a  $\text{Au}_{13}$  cluster precursor: Ligand removal using ozone exposure versus a rapid thermal treatment. *J Catal* 243(1):64–73
103. Bellon P, Manassero M, Sansoni MJ (1972) Crystal and molecular structure of tri-iodoheptakis(tri-*p*-fluorophenylphosphine)undecagold. *Dalton Trans* 1481–1487
104. Bellon PL, Cariati F, Manassero M, Naldini M, Sansoni M (1971) Novel gold clusters. Preparation, properties and x-ray structure determination of salts of octakis(triarylphosphine)enneagold.  $[\text{Au}_9\text{L}_8]\text{X}_3$ . *J Chem Soc Chem Commun* 1423–1424
105. Guliamov O, Frenkel AI, Menard LD, Nuzzo RG, Kronik L (2007) Tangential ligand-induced strain in icosahedral  $\text{Au}_{13}$ . *J Am Chem Soc Commun* 129:10978–10979
106. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. *Phys Rev* 136(3B):B864–B871; Kohn W, Sham L (1965) Self-consistent equations including exchange and correlation effects. *Phys Rev* 140(4A):A1133–A1138
107. Michaelian K, Rendón N, Garzón IL (1999) Structure and energetics of Ni, Ag, and Au nanoclusters. *Phys Rev B* 60(3): 2000–2010
108. Häberlen OD, Chung S, Stener M, Rösch N (1997) From clusters to bulk: A relativistic density functional investigation on a series of gold clusters  $\text{Au}_n$ ,  $n = 6, \dots, 147$ . *J Chem Phys* 106(12):5189–5201
109. Chang CM, Chou MY (2004) Alternative low-symmetry structure for 13-atom metal clusters. *Phys Rev Lett* 93:133401–4
110. Darby S, Mortimer-Jones TV, Johnston RL, Roberts C (2002) Theoretical study of Cu–Au nanoalloy clusters using a genetic algorithm. *J Chem Phys* 116(4):1536–1550
111. Häkkinen H, Barnett RN, Landman U (1999) Electronic structure of passivated  $\text{Au}_{38}(\text{SCH}_3)_{24}$  nanocrystal. *Phys Rev Lett* 82(16):3264–3267
112. Garzon IL, Rovira C, Michaelian K, Beltran MR, Ordejon P, Junquera J, Sanchez-Portal D, Artacho E, Soler JM (2000) Do thiols merely passivate gold nanoclusters? *Phys Rev Lett* 85(24):5250–5251
113. Häkkinen H, Walter M, Grönbeck H (2006) Divide and protect: Capping gold nanoclusters with molecular gold-thiolate rings. *J Phys Chem B* 110(20):9927–9931
114. Dubbois LH, Nuzzo RG (1992) The synthesis, structure, and properties of model organic surfaces. *Annu Rev Phys Chem* 43:437–463
115. Garzón IL, Beltrán MR, González G, Gutierrez-González I, Michaelian K, Reyes-Nava JA, Rodríguez-Hernández JI (2003) Chirality, defects, and disorder in gold clusters. *Eur Phys J D* 24:105–109
116. Knecht MR, Weir MG, Frenkel AI, Crooks RM (2008) Structural rearrangement of bimetallic alloy PdAu nanoparticles to yield core/shell configurations. *Chem Mater* 20:1019–1028
117. Yang JC, Bradley S, Gibson JM (2003) The oblate structure of  $\text{PtRu}_5$  supported on carbon black. *Mater Charact* 51:101–107
118. Wang LL, Khare SV, Johnson DD, Rockett AA, Chirita V, Frenkel AI, Mack NH, Nuzzo RG (2006) Origin of bulk-like structure and bond distributions of  $\text{Pt}_{37}$  and  $\text{Pt}_6\text{Ru}_{31}$  cluster on carbon: comparison of theory and experiment. *J Am Chem Soc* 128(1):131–142
119. Oudenhuijzen MK, Van Bokhoven JA, Miller JT, Ramaker DE, Koningsberger DC (2005) *J Am Chem Soc* 127(5):1530–1540 (and references therein)
120. Wang LL, Johnson DD (2005) Removing critical errors for DFT application to transition-metal nanoclusters: correct ground-state structures of Ru clusters. *J Phys Chem B Lett* 109(49):23113
121. Johnston RL (2003) Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. *Dalton Trans* 4193–4207
122. Kumara S, Johnson DD, Goldberg DE (2007) Scalability of a hybrid extended compact genetic algorithm for ground state optimization of clusters. *Mater Manuf Process* 22(5):570–576
123. Wales DJ, Doye JPK (1997) Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J Phys Chem A* 101(28):5111–5116
124. Wang LL, Johnson DD (2007) Shear instabilities in metallic nanoparticles: Hydrogen-stabilized structure of  $\text{Pt}_{37}$  on carbon. *J Am Chem Soc* 129(12):3658–3664
125. Arslan I, Yates TJV, Browning ND, Midgley PA (2005) Embedded nanostructures revealed in three dimensions. *Science* 309(5744):2195–2198
126. Zuo JM, Vartanyants I, Gao M, Zhang R, Nagahara LA (2003) Atomic resolution imaging of a carbon nanotube from diffraction intensities. *Science* 300(5624):1419–1421
127. Perdew JP, Emzerhof M, Burke K (1996) Rationale for mixing exact exchange with density functional approximations. *J Chem Phys* 105(22):9982–9985
128. Adamo C, Barone V (1999) Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J Chem Phys* 110(13):6158–6170
129. Becke AD (1993) A new mixing of Hartree–Fock and local density-functional theories. *J Chem Phys* 98(2):1372–1377



130. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys* 98(7):5648–5652
131. Bylander DM, Kleinman L (1990) Good semiconductor band gaps with a modified local-density approximation. *Phys Rev B* 41(11):7868–7871
132. Becke AD (1996) Density-functional thermochemistry, vol IV. A new dynamical correlation functional and implications for exact-exchange mixing. *J Chem Phys* 104(3):1040–1046
133. Heyd J, Scuseria GE, Ernzerhof M (2003) Hybrid functionals based on a screened Coulomb potential. *J Chem Phys* 118(18):8207–8215
134. Heyd J, Scuseria GE (2004) Assessment and validation of a screened Coulomb hybrid density functional. *J Chem Phys* 120(16):7274–7280
135. Wu Z, Cohen RE (2006) More accurate generalized gradient approximation for solids. *Phys Rev B* 73(23):235116–235121

### Books and Reviews

- Edelstein AS, Cammarata RC (eds) (1998) *Nanomaterials: Synthesis, Properties and Applications*. Institute of Physics Publishing, Bristol
- Heiz U, Landman U (eds) (2006) *Nanocatalysis*. Springer, Heidelberg

## Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology

RODERICK MELNIK<sup>1,2</sup>

<sup>1</sup> M<sup>2</sup>NeT Lab, Department of Mathematics,  
Wilfrid Laurier University, Waterloo, Canada

<sup>2</sup> Department of Physics, University of Waterloo,  
Waterloo, Canada

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Hierarchy of Mathematical Models for LDSNs](#)

[Numerical Methodologies](#)

[Incorporating New Effects](#)

[Applications and Concluding Remarks](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

#### Low dimensional semiconductor nanostructures

Structures that have characteristic dimensions on the order of nanometers (usually between 1 and 100 nm) and such that the motion of electrons in them can be confined spatially. Quantum well heterostruc-

tures were the first low dimensional semiconductor nanostructures experimentally developed in the early 1970s. Now, a number of different techniques exist to produce a variety of different low dimensional semiconductor nanostructures, including quantum wells, wires, and dots.

**Electronic structure** Electrons move from one energy level to another by emission or absorption of a quantum of energy, a photon. They are placed on electronic orbitals and their configuration defines the electronic structure as the arrangement of electrons in an atom, molecule, etc. The knowledge of the electronic structure of a specific material or a structure helps us in predicting its properties.

**Electromechanical effects** One of the most studied examples of such effects is piezoelectricity, demonstrated for the first time by the brothers Pierre Curie and Jacques Curie in 1880. Now, it is one of the classical examples of coupled phenomena. In the heart of the piezoelectric phenomenon is a coupling mechanism between mechanical and electric fields which is a two-way interaction. In particular, electricity is produced in a piezoelectric body when stress is applied (the direct piezoeffect) and the body is stressed when an electric field is applied (the converse piezoeffect).

**Multiple scales** In studying complex systems we have to deal with coupled phenomena and processes at a multitude of different spatial and temporal scales. Understanding interactions in the system and its response at multiple scales is a fundamental quest of modern science.

**Nanotechnology** A multidisciplinary field that develops and extends our present knowledge into the nanoscale. It is the field where two main science and technology approaches, the “bottom-up” approach and the “top-down” approach, go hand in hand emphasizing the importance of systems science view.

**Mathematical modeling** A universal tool of modern science and technology that uses mathematical language to describe the behavior of systems, processes and phenomena in Nature and man-made.

### Definition of the Subject

Low Dimensional Semiconductor Nanostructures (abbreviated, LDSNs) is a class of physical systems with characteristic dimensions on the order of 1–100 nm such that the motion of their charge carriers can be confined from one, two, or even three spatial dimensions. If we start our consideration from a three dimensional (3D) bulk crystal and create a structure where the motion of carriers is confined



from only one spatial direction, we will arrive at a simplest example of LDSNs, two dimensional ( $2D = 3D - 1D$ ) nanostructures known as quantum wells. Confining the motion of carriers from two spatial directions would lead us to one dimensional ( $1D = 3D - 2D$ ) nanostructures such as quantum wires or quantum rods. Finally, it is possible to confine the motion of charge carriers in LDSNs from all three spatial dimensions. Such nanostructures are often termed as 0-dimensional ( $0D = 3D - 3D$ ) and known in the literature as quantum dots with other terms, such as quantum islands and artificial atoms, also used to describe them.

The history of the subject was closely interwoven first with the works on semiconductor lasers in the early 1970s, but the idea of using heterostructures for emitting purposes can easily be traced back to the early days of electronics and to works of Shockley, Kroemer, Alferov and others [1]. A heterostructure is a semiconductor structure with heterojunction, that is a junction composed of parts (or layers) of dissimilar semiconductor materials that have different electronic properties such as energy (band) gaps. Such structures have properties unmatched by the underlying bulk semiconductor materials. A structure with periodically alternating layers of several materials is called a superlattice. In semiconductor superlattices the bandwidth can be tuned by changing the width of the barriers wells. Such superlattices attracted attention in early 1960s and the early studies of the effect of a periodic modulation of the potential in one direction on the crystal band structure is due to Keldysh. Later, Esaki and Tsu and others used one dimensional models to study transport effects in semiconductor superlattices. In early 1970s, heterostructures were realized in the laboratories.

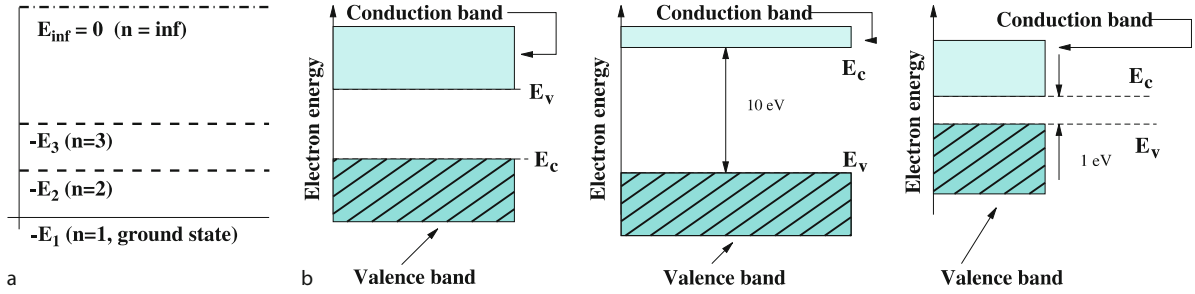
Since that time, the field of low-dimensional systems and nanostructures has grown substantially and includes now semiconductor heterostructures, quantum wells and superlattices, mesoscopic and two-dimensional electron systems, quantum wires and quantum dots. The field has experienced an unprecedented growth in terms of applications, and in addition to its early focus on semiconductor lasers, it includes now a wealth of applications in optoelectronics, quantum information processing, security and defence, health care and biotechnologies, among many other areas.

## Introduction

Echoing R. Feynman's famous comments on the existence of "plenty of room at the bottom", the discussion about nanoscience and complexity continues in both popular and scientific literature [32,68,85]. In the focus of this dis-

cussion is an electron. Electrons have a wave-particle dual nature, occupying regions known as atomic orbitals. Such orbitals form a discrete set of energy levels (see Fig. 1) as they are identified with the quantum states in which electrons, surrounding an atom, may exist. These energy levels (bands) have different widths, depending on the properties of the corresponding orbitals. Mathematically, they are described by a wave function obeying the laws of quantum mechanics, and in particular the Schrödinger equation. In other words, since the electrons of all materials may only have certain allowable energies, we describe each of these allowed energies by the energy levels. From a chemistry point of view, bonding of atoms to form molecules of matter occurs through the interaction of the valence electrons of each atom and in solids atoms are brought together in such a way that the energy levels of individual atoms form bonds of energy.

Electrons tend to occupy energy states with the lowest energy possible and the energy levels corresponding to such states are the valence band's energy levels. The key in differentiating between electric properties of different materials lies with a "forbidden" band (the bandgap), the energy difference that separates the valence band from the more energetic conduction band. More precisely, as seen in Fig. 1, it is the energy difference between the top of the valence band and the bottom of the conduction band. So, from a physics point of view, electrical properties of matter are determined by three main energy bands: valence band, conduction band, and the bandgap. To move into the conduction band, the valence electrons must bridge an energy gap, which determines whether a solid acts as a conductor, a semiconductor, or an insulator (see Fig. 1). Unlike in insulators, in semiconductors enough energy exists in the valence electrons to enable them to cross the energy gap and exist as conduction (free) electrons in the conduction band. It is due to the bandgap that electrons are constrained from jumping easily from the usually more densely populated valence band to the conduction band. Since in a semiconductor bulk, we have continuous energy states, the bandgap is fixed. However, in LDSNs, and in particular in quantum dots, it can be altered to produce a range of energies between the valence and conduction bands, depending on the size, composition, and shape of these structures. Based on the band structures, which can be determined computationally by solving corresponding quantum mechanical models, we can determine optoelectronic properties of LDSNs. These properties are affected by other material properties such as thermal and mechanical, for instance, making the entire problem of determining the properties of LDSNs coupled. This leads to a situation where the systems



**Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology, Figure 1**

**a** A schematic electron energy level diagram; **b** A schematic representation of electric properties of solids, depending on the bandgap: (a) conductors, (b) insulators, (c) semiconductors

science approaches become increasingly important in this field.

Low-dimensional semiconductor nanostructures are multiscale complex systems. Parts of these systems (e. g., two bulk materials) are joined together at the atomic level via interfaces to form a new structure with properties unmatched before [6,16]. As pointed out in [70], “complexity is no longer limited to biology or human sciences: it is invading the physical sciences as deeply rooted in the laws of nature”.

Today, a number of experimental techniques exist to produce both self-assembled as well as free-standing quantum dots. Typical electronic structures of such systems require calculations with  $10^3 - 10^6$  atoms [86]. Already in itself, it is a task of enormous computational complexity. In addition, as applications of nanostructures and nanostructured materials continue to grow rapidly, experimental results clearly point out that there are many additional effects that may influence profoundly optoelectromechanical properties of the nanostructures. Among such effects are strain relaxation, piezoelectric and thermoelectric effects. Indeed, the formation of LDSNs, and in particular quantum dots, is a competition between the surface energy in the structure and strain energy. At the atomistic level, as two different semiconductor materials are joint together, we have a lattice mismatch in the resulting structure that leads to the properties that are substantially different compared to the underlying bulk materials. At the same time, away from the interface between these semiconductor materials, we have to deal with the effects that are pronounced at larger-than-atomistic scales, making the overall problem of determining properties of LDSNs intrinsically multiscale. The solution to this multiscale problem lies with an effective combination of methods based on bottom-up and top-down approaches at the stage of model construction and the application of efficient computational tools at the next stage. We provide further details on these two stages in the next two sections.

### Hierarchy of Mathematical Models for LDSNs

The construction of a hierarchy of mathematical models for modeling LDSNs can be started from the quantum Liouville equation. If the nanostructure has  $n$  degrees of freedom, its state in the Wigner–Weyl phase space can be described by the Wigner distribution function  $\rho$  (the density operator). If we assume that the system’s motion is due to a Hamiltonian function  $H(\mathbf{q}, \mathbf{p}, t)$  [64], then the evolution of the state of the system can be described by the quantum Liouville equation:

$$i\hbar \frac{\partial \rho}{\partial t} = [H, \rho] = H\rho - \rho H, \quad (1)$$

where  $\mathbf{q}$  and  $\mathbf{p}$  are vectors representing the  $n$  coordinate and  $n$  momentum operators, respectively. In a number of applications a generalized form of the quantum Liouville equation has proved to be more convenient where we split the Hamiltonian operator

$$H = \sum_{i=1}^n \frac{p_i^2}{2m_i} + V(\mathbf{q}, t) \quad (2)$$

into two Hamiltonians describing the nuclear motion for the lower and upper electronic states

$$H_l = T + V_l, \quad H_u = T + V_u, \quad (3)$$

where  $T$  is the kinetic energy and  $V_l$  and  $V_u$  are the effective potentials of the nuclei in the lower and upper electronic state, respectively. In this case, the generalized quantum Liouville equation can be given as [13]:

$$i\hbar \frac{\partial \rho}{\partial t} = H_u \rho - \rho H_l, \quad (4)$$

and it can be shown that this equation has the form of a time-dependent Schrödinger equation for a quantum system with  $N = 2n$  degrees of freedom. In what follows

we mainly concern with a non-relativistic time-independent version of this equation:

$$\hat{H}\Psi = E\Psi, \quad (5)$$

where  $E$  is the total energy of (in this general case, many-body) Hamiltonian operator  $\hat{H}$  and  $\Psi$  is the wave function.

A more conventional way to construct a model hierarchy, in particular at the device simulation level, would be to start from the classical Liouville equation for the evolution of the position-velocity probability density  $f(x, v, t)$  and to derive a hierarchy of mathematical models based on the relaxation time approximations, leading to the Boltzmann, hydrodynamic type models and, in the simplest case, to the classical drift-diffusion models [55]. To account for quantum effects in such models several different avenues have been proposed in the literature and examples of such quantum-corrected models include the smooth quantum hydrodynamic approximation and the quantum drift-diffusion model [72]. The resulting models offer a substantial speed up compared to the Schrödinger models. These models start from continuum (e.g., fluid dynamics like) representations and attempt to account for the discrete nature of the problem by incorporating a contribution of quantum effects. This modelling technique is usually associated with the top-down approach. Alternatively, we can start from the Schrödinger type models and move on to incorporate additional important effects that are pronounced at larger-than-atomistic scales. This technique allows us to move from discrete to continuum representations and it is usually associated with the bottom-up approach [56]. As the characteristic dimensions of devices continue to shrink and the number of charge carriers in many applications can be small, highlighting the importance of quantum effects, models based on the mixed-state Schrödinger equation coupled to Poisson's equation for the electrostatic potential have been playing an increasingly important role. However, in the context of LDSNs, such models should be extended to account for strain and the piezoelectric effect responsible for the interaction between electric and mechanical fields, as well as for other coupled effects that may influence the overall properties of the structure.

Atomistic simulation methodologies to solving problem (5) include *ab initio* methods, molecular dynamics and Monte-Carlo. They have a substantial, and often prohibitive for practical applications, computational cost. Inevitably, they are based on a set of assumptions and approximations. One has to approximate the many-body Hamiltonian and the first frequently used simplification lies with the Born–Oppenheimer approximation leading

to decoupling between nuclear and electronic coordinates. This decoupling procedure results in the Schrödinger equation for electrons. In deciding which methodology to apply we should balance between the accuracy of the problem solution and the computational efficiency. For example, quantum dots may contain different amount of charge, from one electrons to thousands [30,38]. If the number of electrons in the structure is small and dynamic properties are not of the major concerns, atomistic methodologies can be preferential. In most realistic cases, however, several other factors should be taken into account. Indeed, if a LDSN is part of a device or a larger structure, the resulting system usually contains a number of atoms far beyond the efficient reach of atomistic methodologies. As a result, the development of other-than-atomistic techniques has become increasingly important. Moreover, even if we are able to complete calculations of the characteristics of the structure with an atomistic methodology, understanding of the interactions between atomistic and larger-than-atomistic scales is often of utmost importance. It has been acknowledged in the literature (e.g., [27,28]) that the key conceptual difficulty in materials science in general, and in computational materials science in particular, is to learn how to deal efficiently with a large range of length and time scales. This is also important for nanostructure studies where accounting for additional phenomena may often bring unexpected results. For instance, for LDSNs nonequilibrium phenomena resulting from fluctuating material interfaces may become of great importance [31].

Even if the time-dependency is neglected in the first approximation in the modelling of LDSNs, as we did by considering model (5), we still have to face a challenge of multiple scales that is different from many other areas of computational materials science. Indeed, for a long period of time, while modelling semiconductor devices (including transport phenomena) by using tools and methods of computational solid state electronics, researchers were able successfully applied classical models where the quantum theory was used for calculating necessary parameters dependent on the microscopic atomistic properties of the structures. As the spatial regions in new devices are becoming comparable with the de Broglie wave length of charged particles, the emphasis on quantum effects is ever increasing. This should come at no surprise because it is exactly the applications of such quantum effects that lead to further advances in nanoelectronics. In the heart of what we call 'quantum devices' is a quantum phenomenon or phenomena where the application of quantum mechanics is essential for predicting properties of such devices. What is often missing in these argu-

ments is the fact that, given passive and contact regions of such devices, in most practical situations they have some “macroscopic” dimensions in a sense that we have to deal with micron or sub-micron regions of the device. In such cases, models that go beyond conventional drift-diffusion approximations and can be applied at sub-micron scales are quite useful [53,54,55]. With an increasing importance of quantum effects, microscopic and macroscopic properties of devices we study in nanotechnological applications are interconnected, they are coupled [56,58]. The study of such properties leads to multiple scale problems in computational materials science that represent non-trivial difficulties for both theory and simulation [56,61]. Relating the quantum mechanical description of the processes and phenomena that are taking place at the atomic level to the behavior of matter at the meso- or macroscopic levels is a key task in studying properties of the materials, the task that has been attempted by many scientists ever since the dawn of quantum mechanics (e. g., [26] and references therein).

### Numerical Methodologies

The band structure calculation is a generic problem occurring in many fields of science and engineering. This problem is not limited to solid state physics applications as we need to face similar problems when we are dealing with wave phenomena in other fields, including acoustics, electromagnetism, and other areas [57]. In addressing this problem in the context LDSNs, the systems science approach becomes most natural. One of the reasons for that has already been emphasized in the previous section – in studying LDSNs we have to deal with intrinsically coupled systems. Another important reason lies with the fact that there is an intrinsic loss of information<sup>1</sup> when we construct a model for the electronic structure calculations. At the level of measurements the problem of the loss of information is well known and widely discussed in the literature (e. g. [8]). In such cases, it is essential to use prior knowledge from other methods and from other scales if the original problem was coarse-grained. Since semiconductor nanostructures are used as parts of electronic and optoelectronic devices, in their modelling a detailed microscopic scale should be combined with larger scales where such effects as strain relaxation and piezoelectric effects are pronounced. A complete description of the system in such cases can be given with a high computational efficiency within the framework of the  $k \cdot p$  theory. It is true that with an increasing computational power (and indeed, we are moving steadily towards a petaflop computer [74]), several

more refined methodologies, including empirical and density functional tight-binding and pseudopotential, become more approachable for the modeling of LDSNs [17]. Nevertheless, it should be realized that the problem remains its multiscale nature as long as the nanostructure under consideration is intended to be embedded as a functional element in a device, a chip, or eventually in a larger scale structure intended for the human use. In such cases, the influence of microscopic and macroscopic domains in the structure should be understood as a two-way interaction. As a result of such interactions, more refined atomistic methodologies mentioned above should be viewed only as a part of the problem solution. We need to integrate effects and phenomena that are pronounced at larger-than-atomistic scales into our mathematical models in such a way that the analysis of the entire system would still be possible within a computationally acceptable time.

Let us consider this issue in more details. It has already been emphasized that in designing new devices, it is paramount to know the properties of the materials they are made of. One of the most important steps in this direction is to find the solution to the electronic and ionic structure problem, achievable only by using efficient computational tools. As we know, under appropriate assumptions the ionic degrees of freedom can be separated from those of the electrons and then the problem of electronic structure calculations can be approached from a number of different directions which rely on a set of underlying assumptions. For example, the problem (5) can be re-formulated by using the variational principle:

$$\delta E[\Psi] = 0, \quad (6)$$

so that if we look for its solutions in the subspace of the products of single particle orbitals we will arrive at the well-known Hartree method. Alternatively, if we seek the solutions to the problem in the subspace of Slater determinants (that is antisymmetrized products of single-particle orbitals), the Hartree–Fock (HF) approximation is obtained. A single Slater determinant of single-electron wave function corresponds to the spin-orbitals. In the latter case, the problem is reduced to  $n$  one-body problems which should be solved self-consistently.

An essentially different concept is used in density functional theory (DFT) approaches where instead of the wave function as in the HF method, we use the spin density (the density of electrons). In practice, by using additional assumptions such as the local density approximation or the generalized gradient approximation we can proceed with the solution of the problem. The problem is usually reduced in this case to the Kohn–Sham (KS) equations requiring a self-consistent solution. From a numerical point

<sup>1</sup>The definition of information can be found in [37].

of view, this can be done in several different ways, e. g. by using the plane-wave cutoff approximation or real-space methodologies [33]. In both cases, the main idea is that the occupied states of the electrons generate a charge density that corresponds to the same potential used to derive the KS equations.

In the context of LDSNs, depending on the specific application at hand, nanostructure modelling can be approached based on several different techniques. The top-down approach provides us a tool to construct a hierarchy of mathematical models of various complexity for semiconductor device modelling [55]. As the size of semiconductor devices becomes smaller, quantum effects, non-homogeneity and defects, along with surface effects, become increasingly important. As a result, an increasing attention is paid to a hierarchy of mathematical models for nanostructures that can be constructed based on the bottom-up approach, starting from the first-principle methods [83]. In a number of cases, this allows us to develop multiscale techniques based on subsequent averaging procedures (e. g., [79]). In modeling LDSNs, it is important to retain the computational efficiency while coupling the scales. Several approaches have been suggested in the literature to address the problem of scale coupling in the context of modelling nanostructures. Classifying multiscale methods, some authors even put the coupling methodologies into a separate group, identifying them with the “direct” coupling methodologies [76]. Note, however, that a coupling procedure is required regardless whether we follow the top-down or bottom-up approach, without formally combining them. A review of several methodologies for the atomistic to continuum coupling can be found in [14]. Other multiscale approaches are based on various forms of averaging and coarse grained techniques [18,80]. In the context of nanostructures with interfaces such techniques have been recently discussed in [40]. Among recently proposed techniques in this direction, we also note [45], where based on the variational multiscale methodology, the authors developed an extension of the two-scale bridging technique to quantum mechanical/continuum coupling. They provided interesting examples, including the coupling procedure between a virtual atom cluster (VAC) model for the continuum modelling with tight binding calculations. In its essence, however, their approach is the top-down approach discussed above. A concurrent multiscale approach that allows one to integrate different scales “seamlessly” was proposed in [46] for metals. Finally, we note that several multiscale techniques based on the renormalization group approach have also recently been discussed [28,31]. The latter approach may prove to be potentially useful for nanostructures if the renormalization

is applied to atomistic models. The underlying idea in the development of all these new methodologies is to achieve “not only higher accuracy, but also more efficient, cost-effective and if possible simpler computational methods in electronic structure calculations” [5,49].

Many applications of LDSNs, and in particular of quantum dots, make the simulations based on atomistic methodologies (ab initio, molecular dynamics or Monte Carlo) impractical. While dealing with an isolated nanostructure we have to deal with at least two scales, the atomic scale at the interfaces and the mesoscopic scale of the structure itself, any such a structure embedded into a working device would require dealing with much more scales. It is this multiscale nature of the problem that makes atomistic approaches very problematic to apply in an efficient manner in applications ranging from basic nanoelectromechanical systems (NEMS) [76] to the development of quantum dot-protein bioconjugate nanoassemblies used for imaging in systems biology [50], or in LDSN applications for RNAi technologies [12]. The field of biological and biomedical applications of LDSNs will continue to grow in the context of the development of new biosensing methodologies [2,36]. By conjugating certain proteins to quantum dots a new class of bioluminescent probes can be created [20]. The models describing these processes and systems are intrinsically multiscale. But even in relatively simple physics applications, the complexity of the problem will increase substantially if we take into account the wetting layer on which the quantum dot is grown. Furthermore, such a systems approach in considering the quantum-dot/wetting layer as a coupled structure will lead to the necessity of the formulation of correct boundary conditions and dealing with an additional disparity in spatial scales between the dot and the wetting layer. This issue of multiple scales and the formulation of correct boundary conditions in this case has been dealt with only recently [58].

In moving to more complicated cases, such as those, for example, discussed above in the context of biological and biomedical applications, an averaging over atomic scales is required. It can be carried out by a number of techniques available, and in particular by the empirical tight-binding, the pseudopotential methodology, or by using the  $k \cdot p$  approximation. As we need to address the problem in a systemic way and to include a range of additional effects that are pronounced at larger-than-atomistic scales, we note that the  $k \cdot p$  theory represents the electronic structure in a continuum-like manner and as such is well suited for incorporating additional effects into the model, including strain relaxation and piezoelectric effects. Within this framework, we can apply a range of pow-



erful numerical discretization procedures well established in mechanics of solids, including finite element methods (FEM). The entire problem can be solved numerically in a computationally efficient manner. It is worthwhile to note further that many methodologies described above, including the KS approach, use a representation of the wave function in a way similar to FEM:

$$\Psi(\mathbf{r}) = \sum_{i=1}^n \alpha_i \psi_i(\mathbf{r}) \quad (7)$$

with  $\{\psi_i\}$  being a set of the basis functions and  $\{\alpha_i\}$  being the set of the coefficients to solve for. For example, models based on a linear combination of atomic orbitals (LCAO) use the representation (7) as does the tight binding model.

Interface boundary conditions in nanostructure modelling deserve special attention. As mentioned, accounting for the correct boundary conditions is often a non-trivial task [58]. Furthermore, when coupling different scales in atomistic and continuum theories, one needs a procedure that should connect these two levels of modelling, atomistic and continuum. Different approaches to construct such procedures have been discussed in the literature. For example, in the context of nanoidentification problems, the authors of [35] surveyed existing so-called “handshake” approaches to coupling the localized fine-grain (e.g., atomistic) domains with their coarse-grain counterparts, where in the latter case the continuum-based modelling is acceptable. Based on the Fourier analysis of the lattice structure, they focused their attention on a particular case of this procedure that allowed them to derive multiscale boundary conditions for the specific problem they analyzed. In the remainder of this section, we show how simple averaging procedures at the fundamental atomistic level allows us to couple quantum mechanical and continuum models for LDSN bandstructure studies in an efficient computational manner. As we have seen, many interesting applications of LDSNs are connected with their optical properties. The main tool for the analysis of the optical properties of these structures is based on atomistic models of the Schrödinger–Poisson (SP) type. However, this model alone is usually not sufficient. Indeed, we recall that the formation of these structures in an industrial setting is based on a competition between the surface energy in the structure and strain energy, making mechanical effects paramount in determining and optimizing the properties of these structures. In addition, coupled effects, such as piezoelectric, could also be important, as it is the case for wurtzite materials (WZ, materials with hexagonal crystal lattice). The piezoeffect, responsible for the two-way coupling between mechanical and electric fields,

is a coupled electromechanical phenomena that is convenient to describe with continuum models [51]. Moreover, as we have recently discussed in [56], nonlinear effects may also become important. As a result, in studying optoelectromechanical properties of low-dimensional semiconductor nanostructures it is essential to combine atomistic and continuum models.

Let us highlight how it can be done. We start from incorporating the lattice mismatch in the models for bandstructure calculations by defining the strain associated with that as a mismatch between two material layers. This is followed by the consideration of the Schrödinger–Poisson model where we account for the piezoelectric effect by coupling the SP model with the model for piezoelectricity. The latter procedure is carried out naturally within the  $k \cdot p$  approximation of the electronic structure. This implies an averaging procedure over the atomistic scales, but as a trade-off, it allows us a straightforward coupling of the atomistic part of the model (Schrödinger–Poisson) with its continuum counterpart (elasticity with piezoeffect). The complete model is then implemented in a finite element (or finite difference) code. Recent results on the influence of electromechanical effects, including piezoelectric and strain contributions, on optoelectronic properties of the structures can be found in [56], where both linear and nonlinear strain components were considered.

The  $k \cdot p$  approximation is based on the first-principle envelope function theory [10,25,39] and offers a computationally attractive tool for simulating LDSNs and the opportunity to incorporate additional effects within its framework. As all techniques arising from the effective mass theory approximations, where an appropriate fitting the model with experimental parameters is required, in the applications of  $k \cdot p$  we may have to deal with the problem of spurious solutions, but several methodologies exist now to overcome this difficulty [40]. In the context of modelling LDSNs, the  $k \cdot p$  framework provides a way for consistently incorporating some of the key larger-scale effects that influence the band structure computed with the SP model. Before proceeding to the description of the main implementation steps of this framework, recall that the accuracy of the  $k \cdot p$  approximation is dependent on the set of basis functions that span the functional space where we seek the envelope function. A practical balance between the physics of the problem and its computational complexity for LDSNs do not usually lead to the choice of “ $n$ ” in (7) higher than 8, which is the case, for example, for wurtzite semiconductors. This leads to models based on the  $8 \times 8$  Hamiltonian. From a physical point of view, the model is based on 6 valence subbands and 2 conduction subbands,

accounting for spin up and down situations and has the form of (5) requiring the solution of the following PDE eigenvalue problem with respect to eigenpair  $(\Psi, E)$ :

$$H\Psi = E\Psi, \quad (8)$$

$$\Psi = (\psi_S^\uparrow, \psi_X^\uparrow, \psi_Y^\uparrow, \psi_Z^\uparrow, \psi_S^\downarrow, \psi_X^\downarrow, \psi_Y^\downarrow, \psi_Z^\downarrow)^T,$$

where  $\psi_X^\uparrow \equiv (|X\rangle \otimes |\uparrow\rangle)$  denotes the wave function component that corresponds to the  $X$  Bloch function of the valence band with the spin function of the missing electron “up”, the subindex “ $S$ ” denotes the wave function component of the conduction band, etc, and  $E$  is the electron/hole energy, as before.

Of course, the form of the Hamiltonian in (8) depends on a particular problem at hand, but a generic representation of the Hamiltonian in the  $k \cdot p$  theory can be given as

$$H \equiv H^{(\alpha, \beta)}(\vec{r}) = -\frac{\hbar^2}{2m_0} \nabla_i \mathcal{H}_{ij}^{(\alpha, \beta)}(\vec{r}) \nabla_j, \quad (9)$$

where  $\mathcal{H}$  is defined by the standard Kohn–Luttinger Hamiltonian or by its refined version based on the Burt–Foreman correction, accounting for the properties of degenerate valence states in an electric field. It represents the kinetic energy plus a nonuniform potential field and other effects contributing to the total potential energy of the system. The superindices  $(\alpha, \beta)$  denote a basis for the wave function of the charge carrier.

Next, we have to incorporate into the model (8), (9) strain relaxation effects in LDSNs. In a sense, accounting for strain effects in this model provides a link between a microscopic (quasi-atomistic after averaging) description of the system with the effects that are pronounced at a larger-than-atomistic scale level as a result of interacting atoms. Indeed, in the case of self-assembled quantum dots, for example, during their growth from the crystal substrate wetting layer, atomic displacements *collectively* induce strain in our finite structure. This fact leads to a modification of the bandstructures obtainable for idealized situations without accounting for strain effects.

Fundamental works in the area by Pikus and Bir, Rashba and Sheka, as well as many others (e. g., [9]) has lead to what is now termed as the Rashba–Sheka–Pikus (RSP) Hamiltonian. Recently, such a Hamiltonian has been applied for the modelling of LDSNs, in particular those made of wurtzite materials (e. g., [23]). The question remains, however, on how to resolve adequately physical effects at edges, corners, and interfaces, including strain nonhomogeneities. Although such effects are quite important in many cases, and in particular for a wide range of quantum dot applications, conventional models for bandstructure calculations are based on the original representa-

tion of strain for the bulk materials [9], where strain can be treated on the basis of infinitesimal theory with the linear Cauchy relationships between strain and displacements. Unfortunately, geometric irregularities can make such approximate models insufficiently accurate.

Another type of nonlinearity that is worthwhile mentioning in the context of modeling LDSNs is related to material nonlinearities. For example, can we still use linear stress-strain relationships in the modelling of LDSN? Since for LDSNs, strain remains orders of magnitudes smaller than their elastic limits, the linear relationship is acceptable, at least in a first approximation. Nevertheless, material nonlinear effects may still be important. Firstly, semiconductors are piezoelectrics and higher order effects may become important at the level of device simulation. Secondly, elastic and dielectric coefficients, being functions of the structure geometry, are nonlinear, but the elasticity in this field is treated by the valence-force-field approaches.

Another very important point is related to the definition of the total potential energy when modelling LDSNs, in particular those made of wurtzite materials. For quite some time, it has been emphasized that both deformational energy and piezoelectric field functionals should be included consistently in the formulation of problems where coupled electromechanical phenomena are pronounced [51,59,66,84]. However, most results obtained so far in the context of bandstructure calculations are pertinent to minimization of elastic energy only [23]. It is true that in some cases, such an approach could be sufficient, in particular for zinc-blende materials (materials with cubic crystal lattice) and where the piezoelectric effect is relatively small, but it does produce inaccurate results in other cases, in particular for wurtzite-based LDSNs. Furthermore, if geometric irregularities are accounted for with nonlinear strain-displacement relationships [62], the resulting coupled models based on the equation of elasticity and the Maxwell equation in dielectric approximation [51,59] become nonlinear. Even in the linear approximation, the coupling between the field of deformation and the piezoelectric field is of fundamental importance [34,84].

One of the main advantages of the  $k \cdot p$  theory lies in its simplicity, but this comes at a cost of being required to fit experimental parameters into the model and generally the number of such parameters increases with the requirement of higher accuracy. For example, the model (8), (9) we described above for wurtzite-based LDSNs would contain 10 parameters. In order to reduce this number, it is important to seek any prior knowledge about the problem such as a specific geometrical shape of LDSN under consideration. Indeed, in the case of cylindrical geometry,

the above model can be reduced to a simpler one. As it has been demonstrated in [41], this can be done by applying the Sercel–Vahala (SV) approach to the Rashba–Sheka–Pikus strain Hamiltonian (9) with details for WZ materials given, e.g., in [23]. As usual, Hamiltonian entries in this case will remain PDE operators, but their representations will be simplified [41]. While the originally discussed model still need to be applied for most quantum dot structures, its simplified version is convenient to use for modelling rods, cylindrical nanowires, as well as many superlattices.

Now, we will highlight the main steps of incorporating coupled physical effects in modelling LDSNs, on the example of strain relaxation and piezoeffect. Firstly, we reduce the Maxwell equation to its dielectric approximation

$$\nabla(\epsilon \nabla \varphi) = -\rho + \nabla \cdot (P^s + P^p) \quad (10)$$

and solve it *simultaneously* with equilibrium equations by using, e.g., the finite element methodology:

$$\frac{\sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} + \frac{\partial \sigma_{xz}}{\partial z} = 0, \quad (11)$$

$$\frac{\sigma_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \sigma_{yz}}{\partial z} = 0, \quad (12)$$

$$\frac{\sigma_{xz}}{\partial x} + \frac{\partial \sigma_{yz}}{\partial y} + \frac{\partial \sigma_{zz}}{\partial z} = 0. \quad (13)$$

These equations are coupled by constitutive stress-strain relationships which differ depending on crystal configuration of the lattice [51,56,59]. In (10)  $P^s$  and  $P^p$  are spontaneous and strain-induced polarization, respectively.

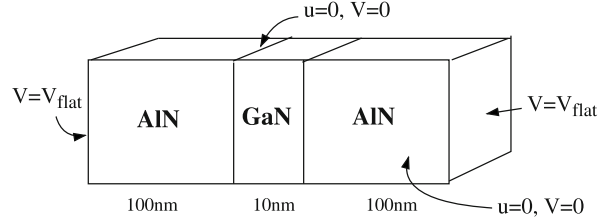
Secondly, the outputs from this model allow us to define the Hamiltonian of system (9) on the same computational grid.

Thirdly, by solving the remaining eight coupled elliptic PDEs (8), we find both eigenfunctions and energies corresponding to all subbands under consideration.

This procedure can be extended to account for the carrier density and charge, in which case an additional coupling loop is necessary between the Schrödinger and Poisson parts of the model [56].

This procedure can be used for modelling most LDSNs, including dots, rods, wires, and superlattices. As a representative example in Fig. 3 we show electronic states in the conduction band in a structure depicted in Fig. 2. Different confinement patterns under coupled and uncoupled situations have important practical implications.

Indeed, LDSNs are designed with an ultimate goal to be used in functional devices where we would like to preserve their properties. For example, in designing new optoelectronic devices, we want to preserve properties of



**Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology, Figure 2**

**A schematic representation of the geometry of the three-layer AlN/GaN WZ nanostructure**

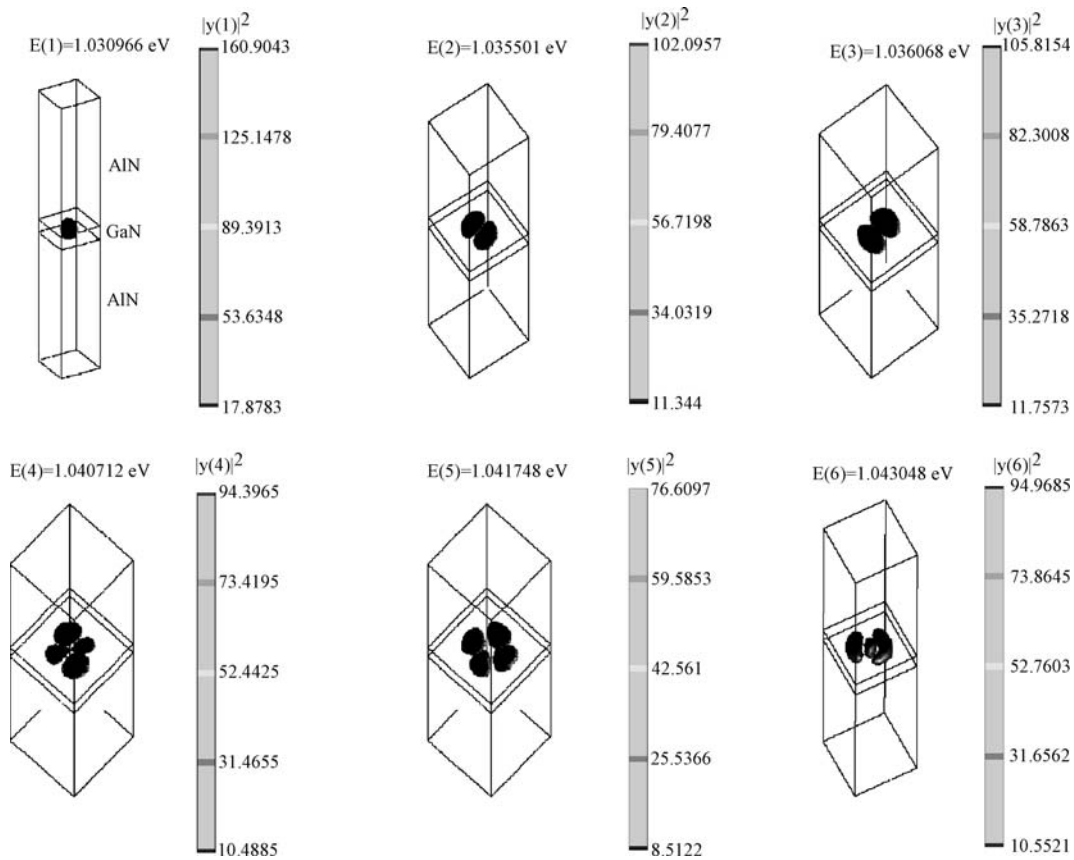
LDSNs such as quantum dots. When such LDSNs are studied theoretically, atomistic (zero-dimensional) density of states in quantum dots allow us to predict some of the characteristics of such structures, e.g. optical gain. If models used for such predictions are uncoupled, a theoretically predicted optical gain can be reduced in practice which may in some cases reduce or even prohibit lasing from the ground state. The reason for that lies with the fact that strain and piezoelectric effects may lead to the elongated ground state carrier function, and hence to a smaller overlap between the electron and hole wave functions, resulting in very different properties as compared to the uncoupled case.

### Incorporating New Effects

In the previous section, we discussed two important effects that can substantially influence electronic bandstructure calculations, strain relaxation and piezoeffects. It has been known for quite some time that we have to deal with an inter-dependence between confinement, strain and piezoelectric effects, but quantitative results with the coupling effects accounted for in predicting nanostructure properties have started appearing relatively recently (e.g., [23,56,66] and references therein). The effects of strain and piezo-charge on the properties of LDSNs have been analyzed within the framework of the envelope function approximation as well as within other approaches discussed in the previous section. As we have already emphasized, a systematic approach is required in incorporating such effects into the existing models. Next, we briefly review other important coupled effects that ultimately should be incorporated in the bandstructure calculation procedures.

### Coupled Phenomena due to Mechanical and Electric Fields Interactions

The piezoeffect is one of the most important examples of electromechanical coupling where mechanical and elec-



**Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology, Figure 3**  
States of conduction band electron confinement (thanks to Roy Mahapatra for producing a figure for this example)

tric fields interact with each other in a two-way interaction manner. The models for coupled dynamic piezoelectricity were put on a rigorous mathematical basis relatively recently [51], and now it is a combination of the piezoelectric effect with semiconductor properties that has attracted attention of the engineers with expectations of a wide range of applications in coupled electronics, sensing and the development of environmentally friendly technologies [82]. Even a new word was coined to reflect these expectations: *nanopiezotronics*. Most of Group III–V compounds used for LDSNs, such as GaAs, have a cubic (zinc-blende) crystal structure and these materials are much better studied. However, it is known also that wide band gap materials such as GaN-based alloys lead to very different properties compared to their zinc-blende counterparts. Usually, they have a hexagonal (wurtzite) crystal structure and a much stronger piezoelectric effect compared to the cubic structures [3]. Anisotropic properties of the material make all the difference, resulting in different properties. The electromechanical coupling is not always linear, and nonlinear electromechanical effects such

as electrostriction have recently attracted the attention of researchers working on the properties of LDSNs and materials [7,29,34,63]. The phenomenon of electromechanical coupling is also in the heart of new types of systems known as quantum electromechanical or nano-electromechanical systems (NEMS). In studying such systems quantum effects also become important, and the other level of coupling comes in this case from quantum mechanics where light is coupled to the matter. The first such coupled model was a model describing the photoelectric effect. A more complex type of coupling is observed in quantum electro-mechanical systems where mechanical motion is quantized with many interesting potential applications reported in the literature (e.g., [78] and references therein).

### Coupled Effects due to Spatio-temporal Interactions in LDSNs

Although acting on very small scales, temporal effects (see p. 518 in [61]) can be important in a number of appli-

cations. These also include many types of NEMS such as quantum shuttle systems, quantum information processing and control. Furthermore, some effects in LDSNs such as Franz–Keldysh oscillations, observed originally in superlattices, require a more careful examination of spatio-temporal interactions in LDSNs. Other spatio-temporal phenomena that have recently been receiving increasing attention are related to phase transformations [60] and phase stability [83] as they have been analyzed in LDSNs [42]. Such phenomena will become even more important for LDSN-based nanocomposites [65] due to the fact that many of their properties which promise new advances in nanotechnological applications are related to interface effects and phenomena.

#### **Coupled Phenomena due to Thermal and Electric Fields Interactions**

Another important group of coupled phenomena important in LDSNs is related to thermoelectricity. Indeed, thermoelectric effects can be substantially enhanced in LDSNs compared to their bulk counterparts [43]. The analysis of such effects is important in the context of the integration of these structures into optoelectronic devices [44]. Furthermore, it is believed that thermoelectric coupling phenomena can facilitate the next generation of electronics and optoelectronics [47].

#### **Coupled Phenomena due to Thermal, Mechanical, and Electric Fields Interactions**

It is not difficult to deduce from the above discussion that a combined effect of thermal, mechanical, and electrical fields should be considered. Models of coupled thermoelectroelasticity are known [52] and await their integration into the models for electronic bandstructure calculations. There is another important reason for incorporating coupled effects due to the thermal field at the nanoscale. Indeed, at the nanoscale junctions of LDSNs, the importance of such effects as local electron heating may also become important to account for [15]. In such cases hydrodynamic types of models are needed in a way similar to what was developed in the past for sub-micron devices [53,54,55]. Such models allow us to deal with non-equilibrium and nonlocal phenomena in an efficient computational manner. Recently, in order to deal with the nonequilibrium quantum many-body problem, the time dependent local deformation approximation theory was proposed allowing to recover both Kohn–Sham and the hydrodynamic formulations [77].

#### **Coupling due to the Magnetic Field**

While it is customary for the modelling of LDSNs to apply the Maxwell equation in its dielectric approximation, there are cases where such an approximation may not be sufficient and the full electromagnetic model could be required. With the magnetic field is accounted for, there are additional nonlinear effects that in many cases would need to be incorporated into the model. For example, it is known that the quantum dot nuclear spin polarization may depend nonlinearly on applied magnetic fields [48].

#### **Many-Body Coupled Effects**

Among such effects are excitonic effects whose importance has been emphasized by a number of authors (e. g., [23] and references therein). Furthermore, in addition to electron-hole Coulomb effects responsible for excitonic transitions, electron-electron Coulomb interactions and the associated Coulomb blockade phenomena, as well as other many-particle effects, may also be important in a number of applications of LDSNs. For example, in the context of quantum dot applications, Coulomb blockade phenomena reflect the finite capacitance of the dot and is effectively an energetic discrimination against two electrons of opposite spin being on the same dot.

#### **More on Nonlinear Coupling**

It is generally true that to better understand the electronic properties of materials with strong correlations between the electrons, nonlinear properties become important. One of the examples is provided by the Kondo effect which is particularly pronounced for the low temperature transport in LDSNs [71]. In addition to nonlinear coupled effects already discussed, in the context of excitonic effects it is important to remember that Rabi oscillations of excitons [75] can show a nonlinear signature in LDSNs such as quantum dot array systems. These oscillations have interesting implications in quantum information processing applications since in the case of a single quantum dot they correspond to the one-qubit rotation (e. g., [75] and references therein). Furthermore, the optical response from LDSNs is typically nonlinear [69].

#### **Spin Coupling**

Spin is the only internal degree of freedom of an electron, and it is natural to attempt to employ the spin-orbit coupling for creating electronic and optoelectronic devices with new functionalities. One of the possibilities includes the injection of nonequilibrium spins with further manipulations of spin polarization at given locations [73].



While coupled systems is a rule rather than an exception in quantum mechanics, spin-orbit coupled systems have a special place among them as they gave birth to the entire new branch of condensed matter physics, spintronics, where we deal with devices whose functionality depends on the manipulation and control of the spin rather than the charge of electrons.

### Applications and Concluding Remarks

As we have seen, LDSNs have led to a number of conceptually new ideas in many applications. Most of these ideas are related, in one way or another, to coupled phenomena being exploited and being put to work in practical applications. By now, strong coupling of light and matter at the single-photon level has led to many important advances [4]. LDSN photonics is an important player in optical information processing [67] and quantum-dot-based photonic crystal lasers represent a reality of today [19]. We are able to create sophisticated nanowires and provide a conceptual framework for nanomachines. In addition to the orbital, we can exploit the spin degree of freedom of the electron and analyze its applicability in spin-based quantum computing, e. g. with the spins of electrons confined to quantum dots [11]. Spin-quantum-NEMS is another example of such an exploitation [22]. We can use LDSNs in energy saving technologies, e. g. in solar panels, as well as in acousto-electronics, photonics, and spintronics with many exciting new developments in these fields happening over the recent years. As the range of applications of LDSNs continue to grow the complex systems science approach in modelling these structures becomes increasingly important. As we have highlighted in the previous sections, in developing new models for LDSNs a systematic combination of bottom-up and top-down techniques is necessary. Based on such a combination a hierarchy of mathematical models of various complexity can be constructed with coupled effects being incorporated in a systematic manner.

Developing this idea, we have discussed two main approaches to nanostructure modelling and highlighted the main procedures for the coupling between quantum mechanical and continuum models. Our approach to multiscale modelling of nanostructures is based on an initial averaging of the quantum mechanical models over the atomic scales and their subsequent coupling to the continuum models. This approach is sufficiently flexible to incorporate other coupled effects discussed in the previous section. We have emphasized that in assisting the design and optimization of new LDSN-based systems by developing the models for LDSNs, it is essential to account for

the coupled effects that may lead to well pronounced modifications of properties of the systems being designed. The modelling framework described here is applicable to both low-density isolated LDSNs as well as dense LDSN arrays. The latter is important, given the fact that many systems we have to deal with can be viewed as complex networks. This also includes quantum networks based on LDSNs with application to quantum information processing [21] and the usage of LDSNs in biological network applications [36]. The area of applications of LDSNs in biology and biomedicine will continue to grow rapidly. Quantum dot-based bioconjugate nanoassembly promises to revolutionize the world we live in. A bridge between biology and nanoscience is being created and such complex systems as DNA-NEMS provides just one, albeit very important, example of this trend. Systems biology and systems science approaches in analyzing such coupled systems are becoming more important than ever and it opens new opportunities for a wider applications of these approaches in this exciting interdisciplinary area.

### Acknowledgments

This work, conducted in the M<sup>2</sup>Net Laboratory (<http://www.m2netlab.wlu.ca>), was supported by the NSERC CRC program, Canada and the Hans Christian Andersen Academy, Denmark. The author is grateful to many his colleagues, in particular to Profs. M. Willatzen, B. Lassen, L. Lew Yan Voon, and R. Mahapatra, for many fruitful discussions on the topics of this paper.

### Bibliography

1. Alferov Z (2000) Double heterostructure lasers: early days and future perspectives. *IEEE J Sel Top Quantum Electron* 6(6): 832–840
2. Alivisatos P (2004) The use of nanocrystals in biological detection. *Nat Biotechnol* 22(1):47–52
3. Andreev AD, O'Reilly EP (2000) Theory of the electronic structure of GaN/AlN hexagonal quantum dots. *Phys Rev* 62(23):15851–15870
4. Aoki T et al (2006) Observation of strong coupling between one atom and a monolithic microresonator. *Nature* 443(12):671–674
5. Artacho E, Beck T, Hernandez E (2006) Current trends in electronic structure: Real-space, embedding and linear scaling techniques. *Phys Status Solidi (b)* 1243(5):971–972
6. Bastard G (1988) *Wave mechanics applied to semiconductor heterostructures*. Halsted Press, New York (a division of John Wiley & Sons)
7. Bester G et al (2006) Effects of linear and nonlinear piezoelectricity on the electronic properties of InAs/GaAs quantum dots. *Phys Rev B* 74:081305(R)
8. Billinge SJ, Levin I (2007) The problem with determining atomic structure at the nanoscale. *Science* 316:561–565

9. Bir GL, Pikus GE (1974) Symmetry and strain-induced effects in semiconductors. Wiley, New York
10. Burt MG (1999) Fundamentals of envelope function theory for electronic states and photonic modes in nanostructures. *J Phys: Condens Matter* 11:R53–R83
11. Cerletti V et al (2005) Recipes for spin-based quantum computing. *Nanotechnology* 16:R27–R49
12. Chen AA et al (2005) Quantum dots to monitor RNAi delivery and improve gene silencing. *Nucleic Acids Res* 33(22):e190
13. Coalson RD, Karplus M (1983) Generalized quantum Liouville equation: its solution by wave packet dynamics. *J Chem Phys* 79(12):6150–6161
14. Curtin WA, Miller RE (2003) Atomistic/continuum coupling in computational materials science. *Model Simul Mater Sci Eng* 11(3):R33–R68
15. D'Agosta R, Sai N, Di Ventra M (2006) Local electron heating in nanoscale conductors. *Nano Lett* 6(12):2935–2938
16. Davies JH (1998) The physics of low-dimensional semiconductors. Cambridge University Press, Cambridge
17. Di Carlo A (2003) Microscopic theory of nanostructured semiconductor devices: beyond the envelope-function approximation. *Semicond Sci Technol* 18:R1–R31
18. E W et al (2007) Heterogeneous multiscale methods: a review. *Commun Comput Phys* 2(3):367–450
19. Ellis B et al (2007) Dynamics of quantum dot photonic crystal lasers. *Appl Phys Lett* 90:151102
20. Evanko D (2006) Bioluminescent quantum dots. *Nat Methods* 3(4):240–241
21. Faraon A et al (2007) Local quantum dot tuning on photonic crystal chips. *Appl Phys Lett* 90:213110
22. Fedorets D et al (2005) Spintronics of a nanoelectromechanical shuttle. *Phys Rev Lett* 95:057203
23. Fonoberov VA, Balandin AA (2003) Excitonic properties of strained wurtzite and zinc-blende  $GaN/Al_xGa_{1-x}N$  quantum dots. *J Appl Phys* 94(11):7178–7186
24. Fonoberov VA, Pokatilov EP, Balandin AA (2003) Interplay of confinement, strain, and piezoelectric effects in the optical spectrum of GaN quantum dots. *J Nanosci Nanotechnol* 3(3):253–256
25. Foreman BA (2005) First-principles envelope-function theory for lattice-matched semiconductor heterostructures. *Phys Rev B (Condens Matter Mater Phys)* 72(16):165345
26. George C, Prigogine I, Rosenfeld L (1972) The macroscopic level of quantum mechanics. *Nature* 240:25–27
27. Ghoniem NM, Busso EP, Kioussis N, Huang H (2003) Multiscale modelling of nanomechanics and micromechanics: an overview. *Philos Mag* 83(31–34):3475–3528
28. Goldenfeld N, Athreya BP, Dantzig JA (2006) Renormalization group approach to multiscale modelling in materials science. *J Stat Phys* 125(5–6):1019–1027
29. Guy IL, Muensit S, Goldys EM (1999) Electrostriction in gallium nitride. *Appl Phys Lett* 75(23):3641–3643
30. Harrison P (2005) Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures. Wiley, New York
31. Haselwandter CA, Vvedensky DD (2007) Multiscale theory of fluctuating interfaces: Renormalization of atomistic models. *Phys Rev Lett* 98(4):046102
32. Heath JR (2001) Nanometer-scale complexity. *Chem Eng News* 79(13):54
33. Heiskanen M et al (2001) Multigrid method for electronic structure calculations. *Phys Rev B* 63:245106
34. Jogai B, Albrecht JD, Pan E (2003) Effect of electromechanical coupling on the strain in AlGaIn/GaN HFETs. *J Appl Phys* 94:3984–3989
35. Karpov E et al (2006) Multiscale boundary conditions in crystalline solids: Theory and application to nanoindentation. *Int J Solids Struct* 43(21):6359–6379
36. Klostianec JM, Chan WCW (2006) Quantum dots in biological and biomedical research: recent progress and present challenges. *Adv Mater* 18:1953–1964
37. Kolmogorov AN (1968) Three approaches to the quantitative definition of information. *Int J Comput Math* 2:157–168
38. Kouwenhoven L, Marcus C (1998) Quantum dots. *Phys World* 6:35–39
39. Lassen B et al (2004) Exact envelope-function theory versus symmetrized Hamiltonian for quantum wires: a comparison. *Solid State Commun* 132(3–4):141–149
40. Lassen B, Melnik RVN, Willatzen M (2007) Spurious solutions in the multiband effective mass theory applied to low dimensional nanostructures. Preprint Series of the Isaac Newton Institute, University of Cambridge
41. Low Yan Voon L, Galieru C, Lassen B et al (2005) Electronic structure of wurtzite quantum dots with cylindrical symmetry. *Appl Phys Lett* 87(4):041906
42. Liang W, Zhou M (2007) Discovery, characterization and modelling of novel shape memory behaviour of fcc metal nanowires. *Philos Mag* 87(14–15):2191–2220
43. Lin Y-M, Dresselhaus MS (2003) Thermoelectric properties of superlattice nanowires. *Phys Rev B* 68:075304
44. Liu W, Balandin AA (2005) Thermoelectric effects in wurtzite GaN and  $Al_xGa_{1-x}N$  alloys. *J Appl Phys* 97:123705
45. Liu WK et al (2006) Bridging scale methods for nanomechanics and materials. *Comp Meth Appl Mech Eng* 195(13–16):1407–1421
46. Lu G, Tadmor EB, Kaxiras E (2006) From electrons to finite elements: a concurrent multiscale approach for metals. *Phys Rev B* 73:024108
47. Majumdar A (2004) Thermoelectricity in semiconductor nanostructures. *Science* 303:777–778
48. Maletinsky P et al (2007) Nonlinear dynamics of quantum dot nuclear spins. *Phys Rev B* 75:035409
49. Martin RM (2004) Electronic structure: basic theory and practical methods. Cambridge University Press, Cambridge
50. Medintz IL et al (2005) Quantum dot bioconjugates for imaging, labelling and sensing. *Nat Mater* 4(6):435–446
51. Melnik RVN (2000) Generalised solutions, discrete models and energy estimates for a 2D problem of coupled field theory. *Appl Math Comput* 107(1):27–55
52. Melnik RVN (2003) Modelling coupled dynamics: Piezoelectric elements under changing temperature conditions. *Int Commun Heat Mass Transf* 30(1):83–92
53. Melnik RVN, He H (2000) Modelling nonlocal processes in semiconductor devices with exponential difference schemes. *J Eng Math* 8(3):233–263
54. Melnik RVN, He H (2000) Quasi-hydrodynamic modelling and computer simulation of coupled thermo-electrical processes in semiconductors. *Math Comput Simul* 52(3–4):273–287
55. Melnik RVN, He H (2000) Relaxation-time approximations of quasi-hydrodynamic type in semiconductor device modelling. *Model Simul Mater Sci Eng* 8(2):133–149

56. Melnik RVN, Mahapatra DR (2007) Coupled effects in quantum dot nanostructures with nonlinear strain and bridging modelling scales. *Comput Struct* 85(11–14):698–711
57. Melnik RVN, Povitsky A (2004) Wave phenomena in physics and engineering: new models, algorithms. *Math Comput Simul* 65(4–5):299–302
58. Melnik RVN, Willatzen M (2004) Bandstructures of conical quantum dots with wetting layers. *Nanotechnology* 15(1):1–8
59. Melnik RVN, Zotsenko KN (2004) Mixed electroelastic waves and CFL stability conditions in computational piezoelectricity. *Appl Numer Math* 48(1):41–62
60. Melnik RVN, Roberts AJ, Thomas KA (2000) Computing dynamics of copper-based SMA via center manifold reduction of 3D models. *Comput Mater Sci* 18:255–268
61. Melnik RVN et al (2003) Distance geometry algorithms in molecular modelling of polymer and composite systems. *Comput Math Appl* 45(1–3):515–534
62. Melnik RVN, Lassen B, Lew Yan Voon LC et al (2005) Nonlinear strain models in the analysis of quantum dot molecules. *Nonlinear Anal* 63(5–7):e2165–e2176
63. Morozovska AN, Eliseev EA, Glinchuk MD (2006) Ferroelectricity enhancement in confined nanorods: direct variational method. *Phys Rev B* 73:214106
64. Narcowich FJ (1986) A Dyson-like expansion for solutions to the quantum Liouville equation. *J Math Phys* 27(10):2502–2510
65. Ovidka IA, Sheinerman AG (2006) Misfit dislocations in nanocomposites with quantum dots, nanowires and their ensembles. *Adv Phys* 55(7–8):627–689
66. Pan E (2002) Elastic and piezoelectric fields around a quantum dot: Fully coupled or semicoupled model? *J Appl Phys* 91(6):3785–3796
67. Pauzauskie PJ, Yang P (2006) Nanowire photonics. *Mater Today* 9(10):36–45
68. Pinto M, Prise KM, Michael BD (2005) Evidence for Complexity at the Nanometer Scale of Radiation-Induced DNA DSBs as a Determinant of Rejoining Kinetics. *Radiat Res* 164(1):73–85
69. Povolotskyi M et al (2004) Non-linear optical properties of InGaAs/AlGaAs nanostructures grown on (N11) surfaces. *Semicond Sci Technol* 19(4):S351–S353
70. Prigogine I (1987) Exploring complexity. *Europ J Oper Res* 30:97–103
71. Pustilnik M (2006) Kondo effect in nanostructures. *Phys Status Solidi (a)* 203(6):1137–1147
72. Radulovic N, Willatzen M, Melnik RVN, Voon LCLY (2006) Influence of the metal contact size on the electron dynamics and transport inside the semiconductor heterostructure nanowire. *J Comput Theor Nanosci* 3(4):551–559
73. Rashba EI (2006) Spin-orbit coupling and spin transport. *Physica E* 34:31–35
74. Sanbonmatsu KY, Tung C-S (2007) High performance computing in biology. : Multimillion atom simulations of nanoscale systems. *J Struct Biol*
75. Slepian GY et al (2004) Rabi oscillations in a semiconductor quantum dot: influence of local fields. *Phys Rev B* 70:045320
76. Tang Z et al (2006) Finite-temperature quasicontinuum method for multiscale analysis of silicon nanostructures. *Phys Rev B* 74:064110
77. Tokatly IV (2007) Time-dependent deformation functional theory. *Phys Rev B* 75:125105
78. Twamley J et al (2006) Spin-detection in a quantum electromechanical shuttle system. *New J Phys* 8:1–27
79. Vvedensky DV (2004) Multiscale modelling of nanostructures. *J Phys Condens Matter* 16:R1537–R1576
80. Wang CY, Zhang X (2006) Multiscale modeling and related hybrid approaches. *Curr Opin Solid State Mater Sci* 10:2–14
81. Wang L et al (2006) Nanoparticles for multiplex diagnostics and imaging. *Nanomedicine* 1(4):413–426
82. Wang ZL (2007) Nanopiezotronics. *Adv Mater* 19:889–892
83. Wen B, Melnik RVN (2008) First principles molecular dynamics study of Cds nanostructure temperature-dependent phase stability. *Appl Phys Lett* 92:261911
84. Willatzen M, Lassen B, Voon LCLY et al (2006) Dynamic coupling of piezoelectric effects, spontaneous polarization, and strain in lattice-mismatched semiconductor quantum-well heterostructures. *J Appl Phys* 100(2):024302
85. Yan H (2004) Nucleic acid nanotechnology. *Science* 306(12):2048–2049
86. Zunger A (1998) Electronic-structure theory of semiconductor quantum dots. *MRS Bull* 2:35–42

---

## Navier–Stokes Equations: A Mathematical Analysis

GIOVANNI P. GALDI

University of Pittsburgh, Pittsburgh, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Derivation of the Navier–Stokes Equations  
and Preliminary Considerations](#)

[Mathematical Analysis  
of the Boundary Value Problem](#)

[Mathematical Analysis  
of the Initial-Boundary Value Problem](#)

[Future Directions](#)

[Acknowledgment](#)

[Bibliography](#)

### Glossary

**Steady-State flow** Flow where both velocity and pressure fields are time-independent.

**Three-dimensional (or 3D) flow** Flow where velocity and pressure fields depend on all three spatial variables.

**Two-dimensional (or planar, or 2D) flow** Flow where velocity and pressure fields depend only on two spa-

tial variables belonging to a portion of a plane, and the component of the velocity orthogonal to that plane is identically zero.

**Local solution** Solution where velocity and pressure fields are known to exist only for a finite interval of time.

**Global solution** Solution where velocity and pressure fields exist for all positive times.

**Regular solution** Solution where velocity and pressure fields satisfy the Navier–Stokes equations and the corresponding initial and boundary conditions in the ordinary sense of differentiation and continuity.

At times, we may interchangeably use the words “flow” and “solution”.

**Basic notation**  $\mathbb{N}$  is the set of positive integers.  $\mathbb{R}$  is the field of real numbers and  $\mathbb{R}^N$ ,  $N \in \mathbb{N}$ , is the set of all  $N$ -tuple  $\mathbf{x} = (x_1, \dots, x_N)$ . The canonical base in  $\mathbb{R}^N$  is denoted by  $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \dots, \mathbf{e}_N\} \equiv \{\mathbf{e}_i\}$ . For  $a, b \in \mathbb{R}$ ,  $b > a$ , we set  $(a, b) = \{x \in \mathbb{R} : a < x < b\}$ ,  $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ ,  $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$  and  $(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$ . By  $\bar{\mathcal{A}}$  we indicate the closure of the subset  $\mathcal{A}$  of  $\mathbb{R}^N$ . A *domain* is an open connected subset of  $\mathbb{R}^N$ . Given a second-order tensor  $\mathbf{A}$  and a vector  $\mathbf{a}$ , of components  $\{A_{ij}\}$  and  $\{a_i\}$ , respectively, in the basis  $\{\mathbf{e}_i\}$ , by  $\mathbf{a} \cdot \mathbf{A}$  [respectively,  $\mathbf{A} \cdot \mathbf{a}$ ] we mean the vector with components  $A_{ij}a_i$  [respectively,  $A_{ij}a_j$ ]. (We use the Einstein summation convention over repeated indices, namely, if an index occurs twice in the same expression, the expression is implicitly summed over all possible values for that index.) Moreover, we set  $|\mathbf{A}| = \sqrt{A_{ij}A_{ij}}$ . If  $\mathbf{h}(\mathbf{z}) \equiv \{h_i(\mathbf{z})\}$  is a vector field, by  $\nabla \mathbf{h}$  we denote the second-order tensor field whose components  $\{\nabla \mathbf{h}\}_{ij}$  in the given basis are given by  $\{\partial h_i / \partial z_j\}$ .

**Function spaces notation** If  $\mathcal{A} \subseteq \mathbb{R}^N$  and  $k \in \mathbb{N} \cup \{0\}$ , by  $C^k(\mathcal{A})$  [respectively,  $C^k(\bar{\mathcal{A}})$ ] we denote the class of functions which are continuous in  $\mathcal{A}$  up to their  $k$ th derivatives included [respectively, are bounded and uniformly continuous in  $\mathcal{A}$  up to their  $k$ th derivatives included]. The subset of  $C^k(\mathcal{A})$  of functions vanishing outside a compact subset of  $\mathcal{A}$  is indicated by  $C_0^k(\mathcal{A})$ . If  $u \in C^k(\mathcal{A})$  for all  $k \in \mathbb{N} \cup \{0\}$ , we shall write  $u \in C^\infty(\mathcal{A})$ . In an analogous way we define  $C^\infty(\bar{\mathcal{A}})$  and  $C_0^\infty(\mathcal{A})$ . The symbols  $L^q(\mathcal{A})$ ,  $W^{m,q}(\mathcal{A})$ ,  $m \geq 0$ ,  $1 \leq q \leq \infty$ , denote the usual Lebesgue and Sobolev spaces, respectively ( $W^{0,q}(\mathcal{A}) = L^q(\mathcal{A})$ ). Norms in  $L^q(\mathcal{A})$  and  $W^{m,q}(\mathcal{A})$  are denoted by  $\|\cdot\|_{q,\mathcal{A}}$ ,  $\|\cdot\|_{m,q,\mathcal{A}}$ . The trace space on the boundary,  $\partial\mathcal{A}$ , of  $\mathcal{A}$  for functions from  $W^{m,q}(\mathcal{A})$  will be denoted by  $W^{m-1/q,q}(\partial\mathcal{A})$  and its norm by  $\|\cdot\|_{m-1/q,q,\partial\mathcal{A}}$ .

By  $D^{k,q}(\mathcal{A})$ ,  $k \geq 1$ ,  $1 < q < \infty$ , we indicate the homogeneous Sobolev space of order  $(m, q)$  on  $\mathcal{A}$ , that is, the class of functions  $u$  that are (Lebesgue) locally integrable in  $\mathcal{A}$  and with  $D^\beta u \in L^q(\mathcal{A})$ ,  $|\beta| = k$ , where  $D^\beta = \partial|\beta|/\partial x_1^{\beta_1} \partial x_2^{\beta_2} \dots \partial x_N^{\beta_N}$ ,  $|\beta| = \beta_1 + \beta_2 + \dots + \beta_N$ . For  $u \in D^{k,q}(\mathcal{A})$ , we put

$$|u|_{k,q,\mathcal{A}} = \left( \sum_{|\beta|=k} \int_{\mathcal{A}} |D^\beta u|^q \right)^{1/q}.$$

Notice that, whenever confusion does not arise, in the integrals we omit the infinitesimal volume or surface elements. Let

$$\mathcal{D}(\mathcal{A}) = \{\boldsymbol{\varphi} \in C_0^\infty(\mathcal{A}) : \operatorname{div} \boldsymbol{\varphi} = 0\}.$$

By  $L_\sigma^q(\mathcal{A})$  we denote the completion of  $\mathcal{D}(\mathcal{A})$  in the norm  $\|\cdot\|_q$ . If  $\mathcal{A}$  is any domain in  $\mathbb{R}^N$  we have  $L^2(\mathcal{A}) = L_\sigma^2(\mathcal{A}) \oplus G(\mathcal{A})$ , where  $G(\mathcal{A}) = \{\mathbf{h} \in L^2(\mathcal{A}) : \mathbf{h} = \nabla p, \text{ for some } p \in D^{1,2}(\mathcal{A})\}$ ; (see Sect. III.1 in [31]). We denote by  $P$  the orthogonal projection operator from  $L^2(\mathcal{A})$  onto  $L_\sigma^2(\mathcal{A})$ . By  $\mathcal{D}_0^{1,2}(\mathcal{A})$  we mean the completion of  $\mathcal{D}(\mathcal{A})$  in the norm  $|\cdot|_{1,2,\mathcal{A}}$ .  $\mathcal{D}_0^{1,2}(\mathcal{A})$  is a Hilbert space with scalar product  $[\mathbf{v}_1, \mathbf{v}_2] := \int_{\mathcal{A}} (\partial \mathbf{v}_1 / \partial x_i) \cdot (\partial \mathbf{v}_2 / \partial x_i)$ . Furthermore,  $\mathcal{D}_0^{-1,2}(\mathcal{A})$  is the dual space of  $\mathcal{D}_0^{1,2}(\mathcal{A})$  and  $\langle \cdot, \cdot \rangle_{\mathcal{A}}$  is the associated duality pairing.

If  $\mathbf{g} \equiv \{g_i\}$  and  $\mathbf{h} \equiv \{h_i\}$  are vector fields on  $\mathcal{A}$ , we set

$$(\mathbf{g}, \mathbf{h})_{\mathcal{A}} = \int_{\mathcal{A}} g_i h_i,$$

whenever the integrals make sense.

In all the above notation, if confusion will not arise, we shall omit the subscript  $\mathcal{A}$ .

Given a Banach space  $X$ , and an open interval  $(a, b)$ , we denote by  $L^q(a, b; X)$  the linear space of (equivalence classes of) functions  $f : (a, b) \rightarrow X$  whose  $X$ -norm is in  $L^q(a, b)$ . Likewise, for  $r$  a non-negative integer and  $I$  a real interval, we denote by  $C^r(I; X)$  the class of continuous functions from  $I$  to  $X$ , which are differentiable in  $I$  up to the order  $r$  included. If  $X$  denotes any space of real functions, we shall use, as a rule, the same symbol  $X$  to denote the corresponding space of vector and tensor-valued functions.

## Definition of the Subject

The Navier–Stokes equations are a mathematical model aimed at describing the motion of an incompressible viscous fluid, like many common ones as, for instance, wa-



ter, glycerin, oil and, under certain circumstances, also air. They were introduced in 1822 by the French engineer Claude Louis Marie Henri Navier and successively re-obtained, by different arguments, by a number of authors including Augustin-Louis Cauchy in 1823, Siméon Denis Poisson in 1829, Adhémar Jean Claude Barré de Saint-Venant in 1837, and, finally, George Gabriel Stokes in 1845. We refer the reader to the beautiful paper by Olivier Darrigol [17], for a detailed and thorough analysis of the history of the Navier–Stokes equations.

Even though, for quite some time, their significance in the applications was not fully recognized, the Navier–Stokes equations are, nowadays, at the foundations of many branches of applied sciences, including Meteorology, Oceanography, Geology, Oil Industry, Airplane, Ship and Car Industries, Biology, and Medicine. In each of the above areas, these equations have collected many undisputed successes, which definitely place them among the most accurate, simple and beautiful models of mathematical physics.

Notwithstanding these successes, up to the present time, a number of unresolved basic mathematical questions remain open – mostly, but not only, for the physically relevant case of three-dimensional flow.

Undoubtedly, the most celebrated is that of proving or disproving the existence of global 3D regular flow for data of arbitrary “size”, no matter how smooth (*global regularity problem*). Since the beginning of the 20th century, this notorious question has challenged several generations of mathematicians who have not been able to furnish a definite answer. In fact, to date, 3D regular flows are known to exist *either* for all times but for data of “small size”, *or* for data of “arbitrary size” but for a finite interval of time only. The problem of global regularity has become so intriguing and compelling that, in the year 2000, it was decided to put a generous bounty on it. In fact, properly formulated, it is listed as one of the seven \$1M Millennium Prize Problems of the Clay Mathematical Institute.

However, the Navier–Stokes equations present also other fundamental open questions. For example, it is not known whether, in the 3D case, the associated initial-boundary value problem is (in an appropriate function space) well-posed in the sense of Hadamard. Stated differently, in 3D it is open the question of whether solutions to this problem exist for all times, are unique and depend continuously upon the data, without being necessarily “regular”.

Another famous, unsettled challenge is whether or not the Navier–Stokes equations are able to provide a rigorous model of turbulent phenomena. These phenomena occur

when the magnitude of the driving mechanism of the fluid motion becomes sufficiently “large”, and, roughly speaking, it consists of flow regimes characterized by chaotic and random property changes for velocity and pressure fields throughout the fluid. They are observed in 3D as well as in two-dimensional (2D) motions (e. g., in flowing soap films). We recall that a 2D motion occurs when the relevant region of flow is contained in a portion of a plane,  $\varpi$ , and the component of the velocity field orthogonal to  $\varpi$  is negligible.

It is worth emphasizing that, in principle, the answers to the above questions may be unrelated. Actually, in the 2D case, the first two problems have long been solved in the affirmative, while the third one remains still open. Nevertheless, there is hope that proving or disproving the first two problems in 3D will require completely fresh and profound ideas that will open new avenues to the understanding of turbulence.

The list of main open problems can not be exhausted without mentioning another outstanding question pertaining to the boundary value problem describing steady-state flow. The latter is characterized by time-independent velocity and pressure fields. In such a case, if the flow region,  $\mathcal{R}$ , is multiply connected, it is not known (neither in 2D nor in 3D) if there exists a solution under a given velocity distribution at the boundary of  $\mathcal{R}$  that *merely* satisfy the physical requirement of conservation of mass.

## Introduction

Fluid mechanics is a very intricate and intriguing discipline of the applied sciences. It is, therefore, not surprising that the mathematics involved in the study of its properties can be, often, extremely complex and difficult. Complexities and difficulties, of course, may be more or less challenging depending on the mathematical model chosen to describe the physical situation.

Among the many mathematical models introduced in the study of fluid mechanics, the Navier–Stokes equations can be considered, without a doubt, the most popular one. However, this does not mean that they can correctly model any fluid under any circumstance. In fact, their range of applicability is restricted to the class of so-called *Newtonian fluids*; see Remark 3 in Sect. “[Derivation of the Navier–Stokes Equations and Preliminary Considerations](#)”. This class includes several common liquids like water, most salt solutions in water, aqueous solutions, some motor oils, most mineral oils, gasoline, and kerosene.

As mentioned in the previous section, these equations were proposed in 1822 by the French engineer Claude



Navier upon the basis of a suitable molecular model. It is interesting to observe, however, that the law of interaction between the molecules postulated by Navier were shortly recognized to be totally inconsistent from the physical point of view for several materials and, in particular, for liquids. It was only more than two decades later, in 1845, that the same equations were re derived by the 26-year-old George Stokes in a quite general way, by means of the theory of continua.

The role of mathematics in the investigation of the fluid properties and, in particular, of the Navier–Stokes equations, is, as in most branches of the applied sciences, twofold and aims at the accomplishment of the following objectives. The first one, of a more fundamental nature, is the validation of the mathematical model, and consists in securing conditions under which the governing equations possess the essential requirements of well-posedness, that is, existence and uniqueness of corresponding solutions and their continuous dependence upon the data. The second one, of a more advanced character, is to prove that the model gives an adequate interpretation of the observed phenomena.

This paper is, essentially, directed toward the first objective. In fact, its goals consist of (1) formulating the primary problems, (2) describing the related known results, and (3) pointing out the remaining basic open questions, for both boundary and initial-boundary value problems. Of course, it seemed unrealistic to give a detailed presentation of such a rich and diversified subject in a few number of pages. Consequently, we had to make a choice which, we hope, will still give a fairly good idea of this fascinating and intriguing subject of applied mathematics.

The plan of this work is the following. In Sect. “[Derivation of the Navier–Stokes Equations and Preliminary Considerations](#)”, we shall first give a brief derivation of the Navier–Stokes equations from continuum theory, then formulate the basic problems and, further on, discuss some basic properties. Section “[Mathematical Analysis of the Boundary Value Problem](#)” is dedicated to the boundary value problem in both bounded and exterior domains. Besides existence and uniqueness, the main topics include the study of the structure of solution set at large Reynolds number, as well as a condensed treatment of bifurcation issues. Section “[Mathematical Analysis of the Initial-Boundary Value Problem](#)” deals with the initial-boundary value problem in a bounded domain and with the related questions of well-posedness and regularity. Finally, in Sect. “[Future Directions](#)” we shall outline some future directions of research. Companion to this last section, is the list of a number of significant open questions that are mentioned throughout the paper.

## Derivation of the Navier–Stokes Equations and Preliminary Considerations

In the continuum mechanics modeling of a fluid,  $\mathcal{F}$ , one assumes that, in the given time interval,  $I \equiv [0, T]$ ,  $T > 0$ , of its motion,  $\mathcal{F}$  continuously fills a region,  $\Omega$ , of the three-dimensional space,  $\mathbb{R}^3$ . We call points, surfaces, and volumes of  $\mathcal{F}$ , *material points* (or *particles*), *material surfaces*, and *material volumes*, respectively. In most relevant applications, the region  $\Omega$  does not depend on time. This happens, in particular, whenever the fluid is bounded by rigid walls like, for instance, in the case of a flow past a rigid obstacle, or a flow in a bounded container with fixed walls. However, there are also some significant situations where  $\Omega$  depends on time as, for example, in the motion of a fluid in a pipe with elastic walls. Throughout this work, we shall consider flow of  $\mathcal{F}$  where  $\Omega$  is *time-independent*.

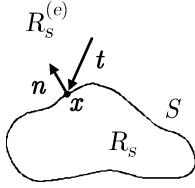
### Balance Laws

In order to describe the motion of  $\mathcal{F}$  it is convenient to represent the relevant physical quantities in the *Eulerian form*. Precisely, if  $\mathbf{x} = (x_1, x_2, x_3)$  is a point in  $\Omega$  and  $t \in [0, T]$ , we let  $\rho = \rho(\mathbf{x}, t)$ ,  $\mathbf{v} = \mathbf{v}(\mathbf{x}, t) = (v_1(\mathbf{x}, t), v_2(\mathbf{x}, t), v_3(\mathbf{x}, t))$  and  $\mathbf{a} = \mathbf{a}(\mathbf{x}, t) = (a_1(\mathbf{x}, t), a_2(\mathbf{x}, t), a_3(\mathbf{x}, t))$  be the density, velocity, and acceleration, respectively, of that particle of  $\mathcal{F}$  that, at the time  $t$ , passes through the point  $\mathbf{x}$ . Furthermore, we denote by  $\mathbf{f} = \mathbf{f}(\mathbf{x}, t) = (f_1(\mathbf{x}, t), f_2(\mathbf{x}, t), f_3(\mathbf{x}, t))$  the external force per unit volume (*body force*) acting on  $\mathcal{F}$ .

In the continuum theory of non-polar fluids, one postulates that in every motion the following equations must hold

$$\left. \begin{aligned} \frac{\partial \rho}{\partial t}(\mathbf{x}, t) + \frac{\partial}{\partial x_i}(\rho(\mathbf{x}, t)v_i(\mathbf{x}, t)) &= 0, \\ \rho(\mathbf{x}, t) a_i(\mathbf{x}, t) &= \frac{\partial T_{ji}}{\partial x_j} + \rho(\mathbf{x}, t) f_i(\mathbf{x}, t), \quad i = 1, 2, 3, \\ T_{ij}(\mathbf{x}, t) &= T_{ji}(\mathbf{x}, t), \quad i, j = 1, 2, 3 \end{aligned} \right\} \begin{array}{l} \text{for all} \\ (\mathbf{x}, t) \in \Omega \\ \times (0, T). \end{array} \quad (1)$$

These equations represent the local form of the *balance laws of  $\mathcal{F}$  in the Eulerian description*. Specifically, (1)<sub>1</sub> expresses the *conservation of mass*, (1)<sub>2</sub> furnishes the *balance of linear momentum*, while (1)<sub>3</sub> is equivalent to the *balance of angular momentum*. The function  $\mathbf{T} = \mathbf{T}(\mathbf{x}, t) = \{T_{ji}(\mathbf{x}, t)\}$  is a second-order, symmetric tensor field, the *Cauchy stress tensor*, that takes into account the *internal forces* exerted by the fluid. More precisely, let  $S$  denote a (sufficiently smooth) fixed, closed surface in  $\mathbb{R}^3$  bounding the region  $\mathcal{R}_S$ , let  $\mathbf{x}$  be any point on  $S$ ,



**Navier–Stokes Equations: A Mathematical Analysis, Figure 1**  
Stress vector at the point  $x$  of the surface  $S$

and let  $\mathbf{n} = \mathbf{n}(\mathbf{x})$  be the outward unit normal to  $S$  at  $\mathbf{x}$ . Furthermore, let  $R_s^{(e)}$  be the region exterior to  $R_s$ , with  $R_s^{(e)} \cap \Omega \neq \emptyset$ . Then the vector  $\mathbf{t} = \mathbf{t}(\mathbf{x}, t)$  defined as

$$\mathbf{t}(\mathbf{x}, t) := \mathbf{n}(\mathbf{x}) \cdot \mathbf{T}(\mathbf{x}, t), \quad (2)$$

represents the force per unit area exerted by the portion of the fluid in  $R_s^{(e)}$  on  $S$  at the point  $\mathbf{x}$  and at time  $t$ ; see Fig. 1.

**Constitutive Equation** An important kinematical quantity associated with the motion of  $\mathcal{F}$  is the *stretching tensor* field,  $\mathbf{D} = \mathbf{D}(\mathbf{x}, t)$ , whose components,  $D_{ij}$ , are defined as follows:

$$D_{ij} = \frac{1}{2} \left( \frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j} \right), \quad i, j = 1, 2, 3. \quad (3)$$

The stretching tensor is, of course, symmetric and, roughly speaking, it describes the rate of deformation of parts of  $\mathcal{F}$ . In fact, it can be shown that a necessary and sufficient condition for a motion of  $\mathcal{F}$  to be *rigid* (namely, the mutual distance between two arbitrary particles of  $\mathcal{F}$  does not change in time), is that  $\mathbf{D}(\mathbf{x}, t) = \mathbf{0}$  at each  $(\mathbf{x}, t) \in \Omega \times I$ . Moreover,  $\operatorname{div} \mathbf{v}(\mathbf{x}, t) \equiv \operatorname{trace} \mathbf{D}(\mathbf{x}, t) \equiv \frac{\partial v_i}{\partial x_i}(\mathbf{x}, t) = 0$  for all  $(\mathbf{x}, t) \in \Omega \times I$ , if and only if the motion is *isochoric*, namely, every material volume does not change with time. A noteworthy class of fluids whose generic motion is isochoric is that of fluids having constant density. In fact, if  $\rho$  is a positive constant, from (1)<sub>1</sub> we find  $\operatorname{div} \mathbf{v}(\mathbf{x}, t) = 0$ , for all  $(\mathbf{x}, t) \in \Omega \times I$ . Fluids with constant density are called *incompressible*. Herein, incompressible fluids will be referred to as *liquids*.

During the generic motion, the internal forces will, in general, produce a deformation of parts of  $\mathcal{F}$ . The relation between internal forces and deformation, namely, the functional relation between  $\mathbf{T}$  and  $\mathbf{D}$ , is called *constitutive equation* and characterizes the physical properties of the fluid. A liquid is said to be *Newtonian* if and only if the relation between  $\mathbf{T}$  and  $\mathbf{D}$  is linear, that is, there exist a scalar function  $p = p(\mathbf{x}, t)$  (the *pressure*) and a constant  $\mu$  (the *shear viscosity*) such that

$$\mathbf{T} = -p\mathbf{I} + 2\mu\mathbf{D}, \quad (4)$$

where  $\mathbf{I}$  denotes the identity matrix. In a *viscous* Newtonian liquid, one assumes that the shear viscosity satisfies the restriction

$$\mu > 0. \quad (5)$$

We will discuss further the meaning of this assumption in Remark 2.

**Navier–Stokes Equations** In view of the condition  $\operatorname{div} \mathbf{v} = 0$ , it easily follows from (4) that

$$\frac{\partial T_{ji}}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \mu \Delta v_i$$

where  $\Delta := \frac{\partial^2}{\partial x_j \partial x_j}$  is the Laplace operator. Therefore, by (1) and (4) we deduce that the equations governing the motion of a Newtonian viscous liquid are furnished by

$$\left. \begin{aligned} \rho a_i &= -\frac{\partial p}{\partial x_i} + \mu \Delta v_i + \rho f_i, \quad i = 1, 2, 3 \\ \frac{\partial v_i}{\partial x_i} &= 0 \end{aligned} \right\} \text{ in } \Omega \times (0, T). \quad (6)$$

It is interesting to observe that *both* equations in (6) are *linear* in all kinematical variables. However, in the Eulerian description, the acceleration is a *nonlinear* functional of the velocity, and we have

$$a_i = \frac{\partial v_i}{\partial t} + v_l \frac{\partial v_i}{\partial x_l},$$

or, in a vector form,

$$\mathbf{a} = \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v}.$$

Replacing this latter expression in (6) we obtain the *Navier–Stokes equations*:

$$\left. \begin{aligned} \rho \left( \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) &= -\nabla p + \mu \Delta \mathbf{v} + \rho \mathbf{f}, \\ \operatorname{div} \mathbf{v} &= 0 \end{aligned} \right\} \text{ in } \Omega \times (0, T). \quad (7)$$

In these equations, the (constant) density  $\rho$ , the shear viscosity  $\mu$  [satisfying (5)], and the force  $\mathbf{f}$  are given quantities, while the *unknowns* are velocity  $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$  and pressure  $p = p(\mathbf{x}, t)$  fields.

Some preliminary comments about the above equations are in order. Actually, we should notice that the unknowns  $\mathbf{v}$ ,  $p$  do not appear in a “symmetric” way. In other words, the equation of conservation of mass (7)<sub>2</sub> does *not* involve the pressure field. This is due to the fact that, from

the mechanical point of view, the pressure plays the role of *reaction force* (Lagrange multiplier) associated with the isochoricity constraint  $\operatorname{div} \mathbf{v} = 0$ . In other words, whenever a portion of the liquid “tries to change its volume” the liquid “reacts” with a suitable distribution of pressure to keep that volume constant. Thus, the pressure field must be generally deduced in terms of the velocity field, once this latter has been determined; see Remark 1.

**Initial-Boundary Value Problem** In order to find solutions to the problem (7), we have to append suitable *initial*, at time  $t = 0$ , and *boundary* conditions. As a matter of fact, these conditions may depend on the specific physical problem we want to model. We shall assume that the region of flow,  $\Omega$ , is bounded by rigid walls,  $\partial\Omega$ , and that the liquid does not “slip” at  $\partial\Omega$ . The appropriate initial and boundary conditions then become, respectively, the following ones

$$\begin{aligned} \mathbf{v}(\mathbf{x}, 0) &= \mathbf{v}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega \\ \mathbf{v}(\mathbf{x}, t) &= \mathbf{v}_1(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \partial\Omega \times (0, T), \end{aligned} \quad (8)$$

where  $\mathbf{v}_0$  and  $\mathbf{v}_1$  are *prescribed* vector fields.

**Steady-State Flow and Boundary-Value Problem** An important, special class of solutions to (7), called *steady-state solutions*, is that where velocity and pressure fields are independent of time. Of course, a necessary requirement for such solutions to exist is that  $\mathbf{f}$  does not depend on time as well. From (7) we thus obtain that a generic steady-state solution,  $(\mathbf{v} = \mathbf{v}(\mathbf{x}), p = p(\mathbf{x}))$ , must satisfy the following equations

$$\left. \begin{aligned} \rho \mathbf{v} \cdot \nabla \mathbf{v} &= -\nabla p + \mu \Delta \mathbf{v} + \rho \mathbf{f} \\ \operatorname{div} \mathbf{v} &= 0 \end{aligned} \right\} \quad \text{in } \Omega. \quad (9)$$

Under the given assumptions on the region of flow, from (8)<sub>2</sub> it follows that the appropriate boundary conditions are

$$\mathbf{v}(\mathbf{x}) = \mathbf{v}_*(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega, \quad (10)$$

where  $\mathbf{v}_*$  is a prescribed vector field.

**Two-Dimensional Flow** In several mathematical questions related to the unique solvability of problems (7)–(8) and (9)–(10), separate attention deserve two-dimensional solutions describing the *planar motions* of  $\mathcal{F}$ . For these solutions the fields  $\mathbf{v}$  and  $p$  depend only on  $x_1, x_2$  (say) [and  $t$  in the case (7)–(8)], and, moreover,  $v_3 \equiv 0$ . Consequently, the relevant (spatial) region of motion,  $\Omega$ , becomes a subset of  $\mathbb{R}^2$ .

**Remark 1** By formally operating with “div” on both sides of (7)<sub>1</sub> and by taking into account (8)<sub>2</sub>, we find that, at each time  $t \in (0, T)$  the pressure field  $p = p(\mathbf{x}, t)$  must satisfy the following Neumann problem

$$\begin{aligned} \Delta p &= \rho(\mathbf{v} \cdot \nabla \mathbf{v} - \mathbf{f}) \quad \text{in } \Omega \\ \frac{\partial p}{\partial n} &= -[\mu \Delta \mathbf{v} - \rho(\mathbf{v}_1 \cdot \nabla \mathbf{v} - \mathbf{f})] \cdot \mathbf{n} \quad \text{at } \partial\Omega \end{aligned} \quad (11)$$

where  $\mathbf{n}$  denotes the outward unit normal to  $\partial\Omega$ . It follows that the *prescription* of the pressure at the bounding walls or at the initial time *independently* of  $\mathbf{v}$ , could be incompatible with (8) and, therefore, could render the problem ill-posed.

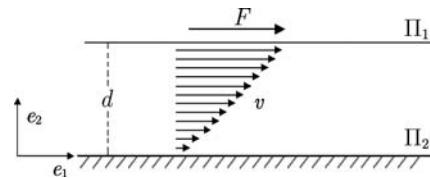
**Remark 2** We can give a simple qualitative explanation of the assumption (5). To this end, let us consider a steady-state flow of a viscous liquid between two parallel, rigid walls  $\Pi_1, \Pi_2$ , set at a distance  $d$  apart and parallel to the plane  $x_2 = 0$  see Fig. 2. The flow is induced by a force per unit area  $\mathbf{F} = F\mathbf{e}_1$ ,  $F > 0$ , applied to the wall  $\Pi_1$ , that moves  $\Pi_1$  with a constant velocity  $\mathbf{V} = V\mathbf{e}_1$ ,  $V > 0$ , while  $\Pi_2$  is kept fixed; see Fig. 2. No external force is acting on the liquid. It is then easily checked that the velocity and pressure fields  $\mathbf{v} = V(x_2/d)\mathbf{e}_1$ ,  $p = p_0 = \text{const.}$  (*pure shear flow*) satisfy (9) with  $\mathbf{f} \equiv \mathbf{0}$ , along with the appropriate boundary conditions  $\mathbf{v}(0) = \mathbf{0}$ ,  $\mathbf{v}(d) = V\mathbf{e}_1$ .

From (2) and (4), we obtain that the force  $\mathbf{t}$  per unit area exerted by the fluid on  $\Pi_1$  is given by

$$\mathbf{t}(x_1, d, x_3) = -\mathbf{e}_2 \cdot \mathbf{T}(x_1, d, x_3) = p_0 \mathbf{e}_2 - \mu \frac{V}{d} \mathbf{e}_1,$$

that is,  $\mathbf{t}$  has a “purely pressure” component  $\mathbf{t}_p = p_0 \mathbf{e}_2$ , and a “purely shear viscous” component  $\mathbf{t}_v = -\mu(V/d)\mathbf{e}_1$ . As expected in a viscous liquid,  $\mathbf{t}_v$  is directed parallel to  $\mathbf{F}$  and, of course, if it is not zero, it should also act *against*  $\mathbf{F}$ , that is,  $\mathbf{t}_v \cdot \mathbf{e}_1 < 0$ . However,  $(V/d) > 0$ , and so we must have  $\mu > 0$ . Since the physical properties of the fluid are independent of the particular flow, this simple reasoning justifies the assumption made in (5).

**Remark 3** As mentioned in the Introduction, the constitutive Eq. (4) and, as a consequence, the Navier–Stokes



**Navier–Stokes Equations: A Mathematical Analysis, Figure 2**  
Pure shear flow between parallel plates induced by a force  $F$

equations, provide a satisfactory model only for a certain class of liquids, while, for others, their predictions are at odds with experimental data. These latter include, for example, biological fluids, like blood or mucus, and aqueous polymeric solutions, like common paints, or even ordinary shampoo. In fact, these liquids, which, in contrast to those modeled by (4), are called *non-Newtonian*, exhibit a number of features that the linear relation (4) is not able to predict. Most notably, in liquids such as blood, the shear viscosity is no longer a constant and, in fact, it decreases as the magnitude of shear (proportional to  $\sqrt{D_{ij}D_{ij}}$ ) increases. Furthermore, liquids like paints or shampoos show, under a given shear rate, a distribution of stress (other than that due to the pressure) in the direction orthogonal to the direction of shear (the so-called *normal stress effect*). Modeling and corresponding mathematical analysis of a variety of non-Newtonian liquids can be found, for example, in [42].

### Mathematical Analysis of the Boundary Value Problem

We begin to analyze the properties of solutions to the boundary-value problem (9)–(10). We shall divide our presentation into two parts, depending on the “geometry” of the region of flow  $\Omega$ . Specifically, we shall treat the cases when  $\Omega$  is either a bounded domain or an exterior domain (flow past an obstacle). For each case we shall describe methods, main results, and fundamental open questions.

#### Flow in Bounded Domains

In this section we shall analyze problem (9)–(10) where  $\Omega$  is a bounded domain of  $\mathbb{R}^3$ . A similar (and simpler) analysis can be performed in the case of planar flow, with exactly the same results.

**Variational Formulation and Weak Solutions** To show existence of solutions, one may use, basically, two types of methods that we shall describe in some detail. The starting point of both methods is the so-called *variational formulation* of (9)–(10). Let  $\varphi \in \mathcal{D}(\Omega)$ . Since  $\operatorname{div} \varphi = \operatorname{div} \mathbf{v} = 0$  in  $\Omega$  and  $\varphi|_{\partial\Omega} = \mathbf{0}$ , we have (by a formal integration by parts)

$$\begin{aligned} (\varphi, \nabla p) &= \int_{\partial\Omega} p \varphi \cdot \mathbf{n} = 0, \\ (\Delta \mathbf{v}, \varphi) &= -[\mathbf{v}, \varphi] + \int_{\partial\Omega} \mathbf{n} \cdot \nabla \mathbf{v} \cdot \varphi = -[\mathbf{v}, \varphi] \\ (\mathbf{v} \cdot \nabla \mathbf{v}, \varphi) &= \int_{\partial\Omega} \mathbf{v} \cdot \mathbf{n} \mathbf{v} \cdot \varphi - (\mathbf{v} \cdot \nabla \varphi, \mathbf{v}) = -(\mathbf{v} \cdot \nabla \varphi, \mathbf{v}) \end{aligned} \quad (12)$$

where  $[\cdot, \cdot]$  is the scalar product in  $\mathcal{D}_0^{1,2}(\Omega)$ . Thus, if we dot-multiply both sides of (9)<sub>1</sub> by  $\varphi \in \mathcal{D}(\Omega)$  and integrate by parts over  $\Omega$ , we (formally) obtain with  $\nu := \mu/\rho$

$$\nu [\mathbf{v}, \varphi] - (\mathbf{v} \cdot \nabla \varphi, \mathbf{v}) = (f, \varphi), \quad \text{for all } \varphi \in \mathcal{D}(\Omega), \quad (13)$$

where we assume that  $f$  belongs to  $\mathcal{D}_0^{-1,2}(\Omega)$ . Equation (13) is the *variational* (or *weak*) form of (9)<sub>1</sub>. We observe that (13) does not contain the pressure. Moreover, every term in (13) is well-defined provided  $\mathbf{v} \in W_{\text{loc}}^{1,2}(\Omega)$ .

**Definition 1** A function  $\mathbf{v} \in W^{1,2}(\Omega)$  is called a weak solution to (9)–(10) if and only if: (i)  $\operatorname{div} \mathbf{v} = 0$  in  $\Omega$ ; (ii)  $\mathbf{v}|_{\partial\Omega} = \mathbf{v}_*$  (in the trace sense); (iii)  $\mathbf{v}$  satisfies (13). If, in particular,  $\mathbf{v}_* \equiv \mathbf{0}$ , we replace conditions (i), (ii) with the single requirement: (i)'  $\mathbf{v} \in \mathcal{D}_0^{1,2}(\Omega)$ .

Throughout this work we shall often use the following result, consequence of the Hölder inequality and of a simple approximating procedure.

**Lemma 1** *The trilinear form*

$$\begin{aligned} (\mathbf{a}, \mathbf{b}, \mathbf{c}) &\in L^q(\Omega) \times W^{1,2}(\Omega) \times L^r(\Omega) \\ &\mapsto (\mathbf{a} \cdot \nabla \mathbf{b}, \mathbf{c}) \in \mathbb{R}, \quad \frac{1}{q} + \frac{1}{r} = \frac{1}{2}, \end{aligned}$$

is continuous. Moreover,  $(\mathbf{a} \cdot \nabla \mathbf{b}, \mathbf{c}) = -(\mathbf{a} \cdot \nabla \mathbf{c}, \mathbf{b})$ , for any  $\mathbf{b}, \mathbf{c} \in \mathcal{D}_0^{1,2}(\Omega)$ , and for any  $\mathbf{a} \in L_\sigma^2(\Omega)$ . Thus, in particular,  $(\mathbf{a} \cdot \nabla \mathbf{b}, \mathbf{b}) = 0$ .

**Regularity of Weak Solutions** If  $f$  is sufficiently regular, the corresponding weak solution is regular as well and, moreover, there exists a scalar function,  $p$ , such that (9) is satisfied in the ordinary sense. Also, if  $\partial\Omega$  and  $\mathbf{v}_*$  are smooth enough, the solution  $(\mathbf{v}, p)$  is smooth up to the boundary and (10) is satisfied in the classical sense. A key tool in the proof of these properties is the following lemma, which is a special case of a more general result due to Cattabriga [12]; (see also Lemma IV.6.2 and Theorem IV.6.1 in [31]).

**Lemma 2** *Let  $\Omega$  be a bounded domain of  $\mathbb{R}^3$ , of class  $C^{m+2}$ ,  $m \geq 0$ , and let  $\mathbf{g} \in \mathbf{W}^{m,q}(\Omega)$ ,  $\mathbf{u}_* \in W^{m+2-1/q,q}(\partial\Omega)$ ,  $1 < q < \infty$ , with  $\int_{\partial\Omega} \mathbf{u}_* \cdot \mathbf{n} = 0$ . Moreover, let  $\mathbf{u} \in W^{1,2}(\Omega)$  satisfy the following conditions*

- (a)  $\nu [\mathbf{u}, \varphi] = (\mathbf{g}, \varphi)$  for all  $\varphi \in \mathcal{D}(\Omega)$ ;
- (b)  $\operatorname{div} \mathbf{u} = 0$ ;
- (c)  $\mathbf{u} = \mathbf{u}_*$  at  $\partial\Omega$  in the trace sense.

*Then,  $\mathbf{u} \in \mathbf{W}^{m+2,q}(\Omega)$  and there exists a unique  $\phi \in W^{m+1,q}(\Omega)$ , with  $\int_{\Omega} \phi = 0$ , such that the pair  $(\mathbf{u}, \phi)$*

satisfies the following Stokes equations

$$\left. \begin{aligned} -\nu \Delta \mathbf{u} &= \nabla \phi + \mathbf{g} \\ \operatorname{div} \mathbf{u} &= 0 \end{aligned} \right\} \quad \text{in } \Omega.$$

Furthermore, there exists a constant  $C = C(\Omega, m, q) > 0$  such that

$$\begin{aligned} \|\mathbf{u}\|_{m+2,q} + \|\phi\|_{m+1,q} \\ \leq C (\|\mathbf{g}\|_{m,q} + \|\mathbf{u}_*\|_{m+2-1/q,q,\partial\Omega}). \end{aligned}$$

To give an idea of how to prove regularity of weak solutions by means of Lemma 2, we consider the case  $\mathbf{f} \in C^\infty(\bar{\Omega})$ ,  $\Omega$  of class  $C^\infty$  and  $\mathbf{v}_* \in C^\infty(\partial\Omega)$ . (For a general regularity theory of weak solutions, see Theorem VIII.52 in [32].) Thus, in particular, by the embedding Theorem II.2.4 in [31]

$$W^{1,2}(\Omega) \subset L^q(\Omega), \quad \text{for all } q \in [1, 6], \quad (14)$$

and by the Hölder inequality we have that  $\mathbf{g} := \mathbf{f} - \mathbf{v} \cdot \nabla \mathbf{v} \in L^{3/2}(\Omega)$ . From (13) and Lemma 2, we then deduce that  $\mathbf{v} \in W^{2,3/2}(\Omega)$  and that there exists a scalar field  $p \in W^{1,3/2}$  such that  $(\mathbf{v}, p)$  satisfy (9) a.e. in  $\Omega$ . Therefore, because of the embedding  $W^{2,3/2}(\Omega) \subset W^{1,3}(\Omega) \subset L^r(\Omega)$ , arbitrary  $r \in [1, \infty)$ , Theorem II.2.4 in [31], we obtain the improved regularity property  $\mathbf{g} \in W^{1,s}(\Omega)$ , for all  $s \in [1, 3/2)$ . Using again Lemma 2, we then deduce  $\mathbf{v} \in W^{3,s}(\Omega)$  and  $p \in W^{2,s}(\Omega)$  which, in particular, gives further regularity for  $\mathbf{g}$ . By induction, we then prove  $\mathbf{v}, p \in C^\infty(\bar{\Omega})$ .

### Existence Results. Homogeneous Boundary Conditions

As we mentioned previously, there are, fundamentally, two kinds of approaches to show existence of weak solutions to (9)–(10), namely, the *finite-dimensional method* and the *function-analytic method*. We will first describe these methods in the case of homogeneous boundary conditions,  $\mathbf{v}_* \equiv \mathbf{0}$ , deferring the study of the non-homogeneous problem to Subsect. “Existence Results. Non-Homogeneous Boundary Conditions”. In what follows, we shall refer to (9)–(10) with  $\mathbf{v}_* = \mathbf{0}$  as (9)–(10)<sub>hom</sub>.

**A. The Finite-Dimensional Method** This popular approach, usually called *Galerkin method*, was introduced by Fujita [28] and, independently, by Vorovich and Yudovich [96]. It consists in projecting (13) on a suitable finite dimensional space,  $V_N$ , of  $\mathcal{D}_0^{1,2}(\Omega)$  and then in finding a solution,  $\mathbf{v}_N \in V_N$ , of the “projected” equation. One then passes to the limit  $N \rightarrow \infty$  to show, with the help of an appropriate uniform estimate, that  $\{\mathbf{v}_N\}$  contains at least one subsequence converging to some

$\mathbf{v} \in \mathcal{D}_0^{1,2}(\Omega)$  that satisfies condition (13). Precisely, let  $\{\boldsymbol{\psi}_k\} \subset \mathcal{D}(\Omega)$  be an orthonormal basis of  $\mathcal{D}_0^{1,2}(\Omega)$ , and set  $\mathbf{v}_N = \sum_{i=1}^N c_{iN} \boldsymbol{\psi}_i$ , where the coefficients  $c_{iN}$  are requested to be solutions of the following nonlinear algebraic system

$$\begin{aligned} \nu[\mathbf{v}_N, \boldsymbol{\psi}_k] - (\mathbf{v}_N \cdot \nabla \boldsymbol{\psi}_k, \mathbf{v}_N) &= (\mathbf{f}, \boldsymbol{\psi}_k), \\ k &= 1, \dots, N. \end{aligned} \quad (15)$$

By means of the Brouwer fixed point theorem, it can be shown (see Lemma VIII.3.2 in [32]) that a solution  $(c_{1N}, \dots, c_{NN})$  to (15) exists, provided the following estimate holds

$$|\mathbf{v}_N|_{1,2} \leq M \quad (16)$$

where  $M$  is a finite, positive quantity independent of  $N$ . Let us show that (16) indeed occurs. Multiplying through both sides of (15) by  $c_{kN}$ , summing over  $k$  from 1 to  $N$ , and observing that, by Lemma 1,  $(\mathbf{v}_N \cdot \nabla \mathbf{v}_N, \mathbf{v}_N) = 0$ , we obtain

$$\nu |\mathbf{v}_N|_{1,2}^2 = (\mathbf{f}, \mathbf{v}_N) \leq |\mathbf{f}|_{-1,2} |\mathbf{v}_N|_{1,2},$$

which proves the desired estimate (16) with  $M := |\mathbf{f}|_{-1,2}$ . Notice that the validity of this latter estimate can be obtained by formally replacing in (13)  $\boldsymbol{\varphi}$  with  $\mathbf{v} \in \mathcal{D}_0^{1,2}(\Omega)$  and by using Lemma 1. From classical properties of Hilbert spaces and from (16), we can select a subsequence  $\{\mathbf{v}_{N'}\}$  and find  $\mathbf{v} \in \mathcal{D}_0^{1,2}(\Omega)$  such that

$$\lim_{N' \rightarrow \infty} [\mathbf{v}_{N'}, \boldsymbol{\varphi}] = [\mathbf{v}, \boldsymbol{\varphi}], \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}_0^{1,2}(\Omega). \quad (17)$$

By Definition 1, to prove that  $\mathbf{v}$  is a weak solution to (9)–(10)<sub>hom</sub> it remains to show that  $\mathbf{v}$  satisfies (13). In view of the *Poincaré inequality* (Theorem II.4.1 in [31]):

$$\|\boldsymbol{\varphi}\|_2 \leq c_P \|\boldsymbol{\varphi}\|_{1,2} \quad \boldsymbol{\varphi} \in \mathcal{D}_0^{1,2}(\Omega), \quad c_P = c_P(\Omega) > 0, \quad (18)$$

by (16) it follows that  $\{\mathbf{v}_N\}$  is bounded in  $W_0^{1,2}(\Omega)$  and so, by Rellich compactness theorem (see Theorem II.4.2 in [31]) we can assume that  $\{\mathbf{v}_{N'}\}$  converges to  $\mathbf{v}$  in  $L^4(\Omega)$ :

$$\lim_{N' \rightarrow \infty} \|\mathbf{v}_{N'} - \mathbf{v}\|_4 = 0. \quad (19)$$

We now consider (15) with  $N = N'$  and pass to the limit  $N' \rightarrow \infty$ . Clearly, by (17), we have for each fixed  $k$

$$\lim_{N' \rightarrow \infty} [\mathbf{v}_{N'}, \boldsymbol{\psi}_k] = [\mathbf{v}, \boldsymbol{\psi}_k]. \quad (20)$$

Moreover, by Lemma 1 and by (19),

$$\lim_{N' \rightarrow \infty} (\mathbf{v}_{N'} \cdot \nabla \boldsymbol{\psi}_k, \mathbf{v}_{N'}) = (\mathbf{v} \cdot \nabla \boldsymbol{\psi}_k, \mathbf{v}),$$



and so, from this latter relation, from (20), and from (15) we conclude that

$$\nu [\mathbf{v}, \boldsymbol{\psi}_k] - (\mathbf{v} \cdot \nabla \boldsymbol{\psi}_k, \mathbf{v}) = \langle \mathbf{f}, \boldsymbol{\psi}_k \rangle, \quad \text{for all } k \in \mathbb{N},$$

Since  $\{\boldsymbol{\psi}_k\}$  is a basis in  $\mathcal{D}_0^{1,2}(\Omega)$ , this latter relation, along Lemma 1, immediately implies that  $\mathbf{v}$  satisfies (13), which completes the existence proof.

**Remark 4** The Galerkin method provides existence of a weak solution corresponding to any given  $\mathbf{f} \in \mathcal{D}_0^{-1,2}(\Omega)$ . Moreover, it is *constructive*, in the sense that the solution can be obtained as the limit of a sequence of “approximate solutions” each of which can be, in principle, evaluated by solving the system of nonlinear algebraic Eqs. (15).

**B. The Function-Analytic Method** This approach, that goes back to the work of Leray [61,63] and of Ladyzhenskaya [58], consists, first, in re-writing (13) as a *nonlinear equation* in the Hilbert space  $\mathcal{D}_0^{1,2}(\Omega)$ , and then using an appropriate topological degree theory to prove existence of weak solutions. Though more complicated than, and not constructive like the Galerkin method, this approach has the advantage of furnishing, as a byproduct, significant information on the solutions set. By (14) and (18) we find

$$\mathcal{D}_0^{1,2}(\Omega) \subset L^4(\Omega), \quad (21)$$

and so, from Lemma 1 and from the Riesz representation theorem, we have that, for each fixed  $\mathbf{v} \in \mathcal{D}_0^{1,2}(\Omega)$ , there exists  $\mathcal{N}(\mathbf{v}) \in \mathcal{D}_0^{1,2}(\Omega)$  such that

$$-(\mathbf{v} \cdot \nabla \boldsymbol{\varphi}, \boldsymbol{\varphi}) = [\mathcal{N}(\mathbf{v}), \boldsymbol{\varphi}], \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}_0^{1,2}(\Omega). \quad (22)$$

Likewise, we have  $\langle \mathbf{f}, \boldsymbol{\varphi} \rangle = [F, \boldsymbol{\varphi}]$ , for some  $F \in \mathcal{D}_0^{1,2}(\Omega)$  and for all  $\boldsymbol{\varphi} \in \mathcal{D}_0^{1,2}(\Omega)$ . Thus, Eq. (13) can be equivalently re-written as

$$N(\mathbf{v}, \mathbf{v}) = F, \quad \text{in } \mathcal{D}_0^{1,2}(\Omega), \quad (23)$$

where the map  $N$  is defined as follows

$$N : (\mathbf{v}, \mathbf{v}) \in (0, \infty) \times \mathcal{D}_0^{1,2}(\Omega) \mapsto \mathbf{v} \mathbf{v} + \mathcal{N}(\mathbf{v}) \in \mathcal{D}_0^{1,2}(\Omega). \quad (24)$$

In order to show the above-mentioned property of solutions, we shall use some basic results related to Fredholm maps of index 0 [84]. This approach is preferable to that originally used by Leray – which applies to maps that are compact perturbations of homeomorphism – because it covers a larger class of problems, including flow in exterior domains [36].

**Definition 2** A map  $M : X \mapsto Y$ ,  $X, Y$  Banach spaces, is Fredholm, if and only if: (i)  $M$  is of class  $C^1$  (in the sense of Fréchet differentiability) and denoted by  $D_x M(x)$  its derivative at  $x \in X$  (iii) the integers. The integers  $\alpha := \dim \{z \in X : [D_x M(x)](z) = 0\}$  and  $\beta := \text{codim} \{y \in Y : [D_x M(x)](z) = y\}$  for some  $z \in X$  are both finite.

The integer  $m := \alpha - \beta$  is independent of the particular  $x \in X$  (Sect. 5.15 in [100]), and is called the *index* of  $M$ .

**Definition 3** A map  $M : X \mapsto Y$ , is said to be proper if  $K_1 := \{x \in X : M(x) = y, y \in K\}$  is compact in  $X$ , whenever  $K$  is compact in  $Y$ .

By using the properties of proper Fredholm maps of index 0 [84], and of the associated *Caccioppoli–Smale degree* [10,84], one can prove the following result (see Theorem I.2.2 in [37])

**Lemma 3** Let  $M$  be a proper Fredholm map of index 0 and of class  $C^2$ , satisfying the following.

- (i) There exists  $\bar{y} \in Y$  such that the equation  $M(x) = \bar{y}$  has one and only one solution  $\bar{x}$ ;
- (ii)  $[D_x M(\bar{x})](z) = 0 \Rightarrow z = 0$ .

Then:

- (a)  $M$  is surjective;
- (b) There exists an open, dense set  $Y_0 \subset Y$  such that for any  $y \in Y_0$  the solution set  $\{x \in X : M(x) = y\}$  is finite and constituted by an odd number,  $\kappa = \kappa(y)$ , of points;
- (c) The integer  $\kappa$  is constant on every connected component of  $Y_0$ .

The next result provides the required functional properties of the map  $N$ .

**Proposition 1** The map  $N$  defined in (24) is of class  $C^\infty$ . Moreover, for any  $\nu > 0$ ,  $N(\nu, \cdot) : \mathcal{D}_0^{1,2}(\Omega) \mapsto \mathcal{D}_0^{1,2}(\Omega)$  is proper and Fredholm of index 0.

*Proof* It is a simple exercise to prove that  $N$  is of class  $C^\infty$  (see Example I.1.6 in [37]). By the compactness of the embedding  $W_0^{1,2}(\Omega) \subset L^4(\Omega)$  Theorem II.4.2 in [31], from Lemma 1 and from the definition of the map  $\mathcal{N}$  [see (22)], one can show that  $\mathcal{N}$  is compact, that is, it maps bounded sequences of  $\mathcal{D}_0^{1,2}(\Omega)$  into relatively compact sequences. Therefore,  $N(\nu, \cdot)$  is a compact perturbation of a multiple of the identity operator. Moreover, by (22) and by Lemma 1 we show that  $[\mathcal{N}(\mathbf{v}), \mathbf{v}] = 0$ , which implies,

$$[N(\nu, \mathbf{v}), \mathbf{v}] = \nu |\mathbf{v}|_{1,2}^2. \quad (25)$$

Using the Schwartz inequality on the left-hand side of (25), we infer that  $|N(v, v)|_{1,2} \geq v|v|_{1,2}$  and so, for each fixed  $v > 0$  we have  $|N(v, v)|_{1,2} \rightarrow \infty$  as  $|v|_{1,2} \rightarrow \infty$ , that is,  $N(v, \cdot)$  is (weakly) *coercive*. Consequently,  $N(v, \cdot)$  is proper (see Theorem 2.7.2 in [7]). Finally, since the derivative of a compact map is compact (see Theorem 2.4.6 in [7]), the derivative map  $D_v N(v, v) \equiv vI + D_v N(v), I$  identity operator, is, for each fixed  $v > 0$ , a compact perturbation of a multiple of the identity, which implies that  $N(v, \cdot)$  is Fredholm of index 0 (see Theorem 5.5.C in [100]).  $\square$

It is easy now to show that, for any fixed  $v > 0$ ,  $N(v, \cdot)$  satisfies all the assumptions of Lemma 3. Actually, in view of Proposition 1, we have only to prove the validity of (i) and (ii). We take  $\bar{y} \equiv 0$ , and so, from (25) we obtain that the equation  $N(v, v) = 0$  has only the solution  $\bar{x} \equiv v = 0$ , so that (i) is satisfied. Furthermore, from (22) it follows that the equation  $D_v N(v, 0)(w) = 0$  is equivalent to  $v w = 0$ , and so also condition (ii) is satisfied for  $v > 0$ . Thus, we have proved the following result (see also [22]).

**Theorem 1** For any  $v > 0$  and  $F \in \mathcal{D}_0^{1,2}(\Omega)$ , Eq. (23) has at least one solution  $v \in \mathcal{D}_0^{1,2}(\Omega)$ . Moreover, for each fixed  $v > 0$ , there exists open and dense  $\mathcal{O} = \mathcal{O}(v) \subset \mathcal{D}_0^{1,2}(\Omega)$  with the following properties: (i) For any  $F \in \mathcal{O}$  the number of solutions to (23),  $n = n(F, v)$  is finite and odd; (ii) the integer  $n$  is constant on each connected component of  $\mathcal{O}$ .

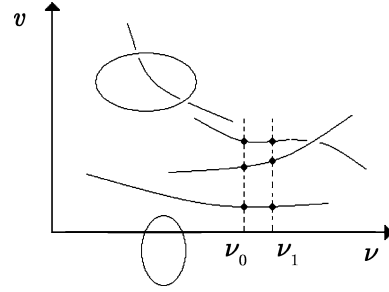
Next, for a given  $F \in \mathcal{D}_0^{1,2}(\Omega)$ , consider the solution manifold

$$S(F) = \{(v, v) \in (0, \infty) \times \mathcal{D}_0^{1,2}(\Omega) : N(v, v) = F\}.$$

By arguments similar to those used in the proof of Theorem 1, one can show the following “generic” characterization of  $S(F)$  see Example I.2.4 in [37] and also Chap. 10.3 in [93].

**Theorem 2** There exists dense  $\mathcal{P} \subset \mathcal{D}_0^{1,2}(\Omega)$  such that, for every  $F \in \mathcal{P}$ , the set  $S(F)$  is a  $C^\infty$  1-dimensional manifold. Moreover, there exists an open and dense subset of  $(0, \infty)$ ,  $\Lambda = \Lambda(F)$ , such that for each  $v \in \Lambda$ , Eq. (23) has a finite number  $m = m(F, v) > 0$  of solutions. Finally, the integer  $m$  is constant on every open interval contained in  $\Lambda$ .

In other words, Theorem 2 expresses the property that, for every  $F \in \mathcal{P}$ , the set  $S(F)$  is the union of smooth and non-intersecting curves. Furthermore, “almost all” lines  $v = v_0 = \text{const.}$  intersect these curves at a finite number of points,  $m(v_0, F)$ , each of which is a solution to (23) corresponding to  $v_0$  and to  $F$ . Finally,  $m(v_0, F) = m(v_1, F)$



Navier–Stokes Equations: A Mathematical Analysis, Figure 3  
Sketch of the manifold  $S(F)$

whenever  $v_0$  and  $v_1$  belong to an open interval of a suitable dense set of  $(0, \infty)$ . A sketch of the manifold  $S(F)$  is provided in Fig. 3.

**Existence Results. Non-Homogeneous Boundary Conditions** Existence of solutions to (9)–(10) when  $v_* \neq 0$  leads to one of the most challenging open questions in the mathematical theory of the Navier–Stokes equations, in the case when the boundary  $\partial\Omega$  is constituted by more than one connected component. In order to explain the problem, let  $S_i$ ,  $i = 1, \dots, K$ ,  $K \geq 1$ , denote these components. Conservation of mass (9)<sub>2</sub> along with Gauss theorem imply the following *compatibility condition* on the data  $v_*$

$$\sum_{i=1}^K \int_{S_i} v_* \cdot n_i \equiv \sum_{i=1}^K \Phi_i = 0, \quad (26)$$

where  $n_i$  denotes the outward unit normal to  $S_i$ . From the physical point of view, the quantity  $\rho \Phi_i$  represents the mass flow-rate of the liquid through the surface  $S_i$ . Now, assuming  $v_*$  and  $\Omega$  sufficiently smooth (for example,  $v_* \in W^{1/2,2}(\partial\Omega)$  and  $\Omega$  locally Lipschitzian Sect. VIII.4 in [32]), we look for a weak solution to (9)–(10) in the form  $v = u + V$ , where  $V \in W^{1,2}(\Omega)$  is an extension of  $v_*$  with  $\text{div } V = 0$  in  $\Omega$ . Thus, if we use, for example, the Galerkin method of Subsect. IV.1.1(A), from (15) we obtain that the “approximate solution”  $u_N = \sum_{i=1}^N c_i N \psi_i$  must satisfy the following equations ( $k = 1, \dots, N$ )

$$\begin{aligned} & v[u_N, \psi_k] - (u_N \cdot \nabla \psi_k, u_N) - (u_N \cdot \nabla \psi_k, V) \\ & - (V \cdot \nabla \psi_k, u_N) - (V \cdot \nabla \psi_k, V) + v[V, \psi_k] \\ & = \langle f, \psi_k \rangle. \end{aligned} \quad (27)$$

Therefore, existence of a weak solution will be secured provided we show that the sequence  $\{v_N := u_N + V\}$  satisfies the bound (16). In turn, this latter is equivalent to

showing

$$|\mathbf{u}_N|_{1,2} \leq M_1, \quad (28)$$

where  $M_1$  is a finite, positive quantity *independent of*  $N$ . Multiplying through both sides of (27) by  $c_{kN}$ , summing over  $k$  from 1 to  $N$ , and observing that, by Lemma 1,  $(\mathbf{u}_N \cdot \nabla \mathbf{u}_N, \mathbf{u}_N) = (\mathbf{V} \cdot \nabla \mathbf{u}_N, \mathbf{u}_N) = 0$ , we find

$$\begin{aligned} & \nu |\mathbf{u}_N|_{1,2}^2 \\ &= (\mathbf{u}_N \cdot \nabla \mathbf{u}_N, \mathbf{V}) + (\mathbf{V} \cdot \nabla \mathbf{u}_N, \mathbf{V}) - [\mathbf{V}, \mathbf{u}_N] + \langle \mathbf{f}, \mathbf{u}_N \rangle. \end{aligned}$$

By using (14), the Hölder inequality and the Cauchy–Schwartz inequality

$$ab \leq \varepsilon a^2 + (1/4\varepsilon)b^2, \quad a, b, \varepsilon > 0 \quad (29)$$

on the last three terms on the right-hand side of this latter equation, we easily find, for a suitable choice of  $\varepsilon$ ,

$$\frac{\nu}{2} |\mathbf{u}_N|_{1,2}^2 \leq (\mathbf{u}_N \cdot \nabla \mathbf{u}_N, \mathbf{V}) + C, \quad (30)$$

where  $C = C(\mathbf{V}, \mathbf{f}, \Omega) > 0$ . From (30) it follows that, in order to obtain the bound (28) *without restrictions on the magnitude of*  $\nu$ , it suffices that  $\mathbf{V}$  meets the following requirement:

$$\begin{aligned} & \text{Given } \varepsilon, \quad \text{there is } \mathbf{V} = \mathbf{V}(\varepsilon, \mathbf{x}) \in (0, \nu/4) \\ & \text{such that } (\boldsymbol{\varphi} \cdot \nabla \boldsymbol{\varphi}, \mathbf{V}) \leq \varepsilon |\boldsymbol{\varphi}|_{1,2}^2 \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega). \end{aligned} \quad (31)$$

As indicated by Leray pp. 28–30 in [61] and clarified by Hopf [52], if  $\Omega$  is smooth enough and if  $K = 1$ , that is, if  $\partial\Omega$  is constituted by only one connected component,  $S_1$ , it is possible to construct a family of extensions satisfying (31). Notice that, in such a case, condition (26) reduces to the single condition  $\Phi_1 = 0$ . If  $K > 1$ , the same construction is still possible but *with the limitation that*  $\Phi_i = 0$ , *for all*  $i = 1, \dots, K$ . It should be emphasized that this condition is quite restrictive from the physical point of view, in that it does not allow for the presence, in the region of flow, of isolated “sources” and “sinks” of liquid. Nevertheless, one may wonder if, by using a different construction, it is still possible to satisfy (31). Unfortunately, as shown by Takeshita [90] by means of explicit examples, *in general, the existence of extensions satisfying (31) implies*  $\Phi_i = 0$ , *for all*  $i = 1, \dots, K$ ; see also § VIII.4 in [32].

We wish to emphasize that the same type of conclusion holds if, instead of the Galerkin method, we use the function-analytic approach; see [61], notes to Chap. VIII in [32].

Finally, it should be remarked that, in the special case of *two-dimensional* domains possessing suitable symmetry and of symmetric boundary data, Amick [1] and Fujita [29] have shown existence of corresponding symmetric solutions under the general assumption (26). However, we have the following.

**Open Question** Let  $\Omega$  be a smooth domain in  $\mathbb{R}^n$ ,  $n = 2, 3$ , with  $\partial\Omega$  constituted by  $K > 1$  connected components,  $S_i$ ,  $i = 1, \dots, K$ , and let  $\mathbf{v}_*$  be any smooth field satisfying (26). *It is not known if the corresponding problem (9)–(10) has at least one solution.*

**Uniqueness and Steady Bifurcation** It is a well-established experimental fact that a steady flow of a viscous incompressible liquid is “observed”, namely, it is unique and stable, if the magnitude of the driving force, usually measured through a dimensionless number  $\lambda \in (0, \infty)$ , say, is below a certain threshold,  $\lambda_c$ . However, if  $\lambda > \lambda_c$ , this flow becomes unstable and another, different flow is instead observed. This latter may be steady or unsteady (typically, time-periodic). In the former case, we say that a *steady bifurcation* phenomenon has occurred. From the physical point of view, bifurcation happens because the liquid finds a more “convenient” motion (than the original one) to dissipate the increasing energy pumped in by the driving force. From the mathematical point of view, bifurcation occurs when, roughly speaking, two solution-curves (parametrized with the appropriate dimensionless number) intersect. (As we know from Theorem 2, the intersection of these curves is not “generic”.) It can happen that one curve exists for all values of  $\lambda$ , while the other only exists for  $\lambda > \lambda_c$  (*supercritical bifurcation*). The point of intersection of the two curves is called the *bifurcation point*. Thus, in any neighborhood of the bifurcation point, we must have (at least) two distinct solutions and so a necessary condition for bifurcation is the occurrence of non-uniqueness. This section is dedicated to the above issues.

**Uniqueness Results** It is simple to show that, if  $|\mathbf{v}|_{1,2}$  is not “too large”, then  $\mathbf{v}$  is the only weak solution to (9)–(10) corresponding to the given data.

**Theorem 3 (Uniqueness)** *Let  $\Omega$  be locally Lipschitzian and let*

$$|\mathbf{v}|_{1,2} < \nu/\kappa, \quad (32)$$

*with  $\kappa = \kappa(\Omega) > 0$ . Then, there is only one weak solution to (9)–(10).*

*Proof* Let  $\mathbf{v}, \mathbf{v}_1 = \mathbf{v} + \mathbf{u}$  be two different solutions. From (13) we deduce

$$\nu[\mathbf{u}, \boldsymbol{\varphi}] = (\mathbf{u} \cdot \nabla \boldsymbol{\varphi}, \mathbf{v}) + (\mathbf{v}_1 \cdot \nabla \boldsymbol{\varphi}, \mathbf{u}), \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega). \quad (33)$$

If  $\Omega$  is locally Lipschitzian, then  $\mathbf{u} \in \mathcal{D}_0^{1,2}(\Omega)$  (see Theorem II.3.2 and § II.3.5 in [31]), and since  $\mathcal{D}(\Omega)$  is dense in  $\mathcal{D}_0^{1,2}(\Omega)$ , with the help of Lemma 1 we can replace  $\boldsymbol{\varphi}$  with  $\mathbf{u}$  in (33) to get

$$\nu|\mathbf{u}|_{1,2}^2 = (\mathbf{u} \cdot \nabla \mathbf{u}, \mathbf{v}).$$

We now use (14) and Lemma 1 on the right-hand side of this equation to find

$$(\nu - \kappa |\mathbf{v}|_{1,2})|\mathbf{u}|_{1,2}^2 \leq 0,$$

with  $\kappa = \kappa(\Omega) > 0$ , which proves  $\mathbf{u} = 0$  in  $\mathcal{D}_0^{1,2}(\Omega)$ , namely, uniqueness, if  $\mathbf{v}$  satisfies (33).  $\square$

**Remark 5** As we have seen previously, if  $\mathbf{v}_* \equiv \mathbf{0}$ , weak solutions satisfy  $\nu|\mathbf{v}|_{1,2} \leq |\mathbf{f}|_{-1,2}$ . Thus, (32) holds if  $|\mathbf{f}|_{-1,2} \leq \nu^2/\kappa$ . If  $\mathbf{v}_* \neq \mathbf{0}$ , then one can show that (32) holds if  $|\mathbf{f}|_{-1,2} + (1 + \nu)\|\mathbf{v}_*\|_{1/2,2,\partial\Omega} + \|\mathbf{v}_*\|_{1/2,2,\partial\Omega}^2 \leq \nu^2/\kappa_1$ , where  $\kappa_1$  has the same properties as  $\kappa$ ; (see Theorem VIII.4.2 in [32]). Notice that these conditions are satisfied if a suitable non-dimensional parameter  $\lambda \sim (|\mathbf{f}|_{-1,2} + \|\mathbf{v}_*\|_{1/2,2,\partial\Omega})$  is “sufficiently small”.

Remarkably enough, one can give *explicit examples of non-uniqueness*, if condition (32) is violated. More specifically, we have the following result (see Theorem VIII.2.2 in [32]).

**Theorem 4 (Non-Uniqueness)** *Let  $\Omega$  be a bounded smooth body of revolution around an axis  $r$ , that does not include points of  $r$ . For example,  $\Omega$  is a torus of arbitrary bounded smooth section. Then there are smooth fields  $\mathbf{f}$  and  $\mathbf{v}_*$  and a value of  $\nu > 0$  such that problem (9)–(10) corresponding to these data admits at least two distinct and smooth solutions.*

**Some Bifurcation Results** By using the functional setting introduced in Sect. “B. The Function-Analytic Method”, it is not difficult to show that steady bifurcation can be reduced to the study of a suitable nonlinear eigenvalue problem in the space  $\mathcal{D}_0^{1,2}(\Omega)$ . To this end, we recall certain basic definitions.

Let  $U$  be an open interval of  $\mathbb{R}$  and let

$$M : (x, \mu) \in X \times U \mapsto Y. \quad (34)$$

**Definition 4** The point  $(x_0, \mu_0)$  is called a bifurcation point of the equation

$$M(x, \mu) = 0 \quad (35)$$

if and only if (a)  $M(x_0, \mu_0) = 0$ , and (b) there are (at least) two sequences of solutions,  $\{(x_m, \mu_m)\}$  and  $\{(x_m^*, \mu_m^*)\}$ , to (35), with  $x_m \neq x_m^*$ , for all  $m \in \mathbb{N}$ , such that  $(x_m, \mu_m) \rightarrow (x_0, \mu_0)$  and  $(x_m^*, \mu_m^*) \rightarrow (x_0, \mu_0)$  as  $m \rightarrow \infty$ .

If  $M$  is suitably smooth around  $(x_0, \mu_0)$ , a necessary condition for  $(x_0, \mu_0)$  to be a bifurcation point is that  $D_x M(x_0, \mu_0)$  is not a bijection. In fact, we have the following result which is an immediate corollary of the implicit function theorem (see, e.g., Lemma III.1.1 in [37]).

**Lemma 4** *Suppose that  $D_x M$  exists in a neighborhood of  $(x_0, \mu_0)$ , and that both  $M$  and  $D_x M$  are continuous at  $(x_0, \mu_0)$ . Then, if  $(x_0, \mu_0)$  is a bifurcation point of (34),  $D_x M(x_0, \mu_0)$  is not a bijection. If, in particular,  $D_x M(x_0, \mu_0)$  is a Fredholm operator of index 0 (see Definition 2), then the equation  $D_x M(x_0, \mu_0)x = 0$  has at least one nonzero solutions.*

Let  $\mathbf{v}_0 = \mathbf{v}_0(\nu)$ ,  $\nu \in (0, \infty)$  be a family of weak solution to (9)–(10) corresponding to given data  $\mathbf{f}$  and  $\mathbf{v}_*$ . We assume that  $\mathbf{f}$  and  $\mathbf{v}_*$  are fixed. Denoting by  $\mathbf{v}$  any other solution corresponding to the same data, by an argument completely similar to that leading to (23), we find that  $\mathbf{u} := \mathbf{v} - \mathbf{v}_0$ , satisfies the following equation in  $\mathcal{D}_0^{1,2}(\Omega)$

$$\nu \mathbf{u} + \mathbf{B}(\mathbf{v}_0)(\mathbf{u}) + \mathcal{N}(\mathbf{u}) = \mathbf{0}, \quad (36)$$

with  $\mathbf{B}(\mathbf{v}_0) := D_v \mathcal{N}(\mathbf{v}_0)$  and  $\mathcal{N}$  defined in (22). Obviously,  $(\mathbf{v}_0(\nu_0), \nu_0)$  is a bifurcation point for the original equation if and only if  $(\mathbf{0}, \nu_0)$  is such for (36). Thus, in view of Lemma 4, a necessary condition for  $(\mathbf{0}, \nu_0)$  to be a bifurcation point for (36) is that the equation

$$\nu_0 \mathbf{v} + \mathbf{B}(\mathbf{v}_0(\nu_0))(\mathbf{v}) = \mathbf{0} \quad (37)$$

has at least one nonzero solution  $\mathbf{v} \in \mathcal{D}_0^{1,2}(\Omega)$ .

In several significant situations it happens that, after a suitable non-dimensionalization of (36), the family of solutions  $\mathbf{v}_0(\nu)$ ,  $\nu \in (0, \infty)$  is independent of the parameter  $\nu$  which, this time, has to be interpreted as the inverse of an appropriate dimensionless number (*Reynolds number*), like, for instance, in the Taylor–Couette problem; see the following subsection. Now, from Proposition 1, we know that  $\mathbf{B}(\mathbf{u})$  is compact at each  $\mathbf{u} \in \mathcal{D}_0^{1,2}(\Omega)$ , so that  $\nu \mathbf{I} + \mathbf{B}(\mathbf{u})$  is Fredholm of index 0, at each  $\mathbf{u} \in \mathcal{D}_0^{1,2}(\Omega)$ , for all  $\nu > 0$  (see Theorem 5.5.C in [100]). Therefore, whenever  $\mathbf{v}_0$  does not depend on  $\nu$ , in a neighborhood of

$v_0$ , from Lemma 4 we find that a *necessary condition* for  $(v_0, v_0)$  to be a bifurcation point for (35) is that  $v_0$  is an *eigenvalue of the (compact) linear operator  $B(v_0)$* .

The stated condition becomes also *sufficient*, provided we make the additional assumptions that  $v_0$  is a *simple eigenvalue* of  $B(v_0)$ . This is a consequence of the following theorem (see, e. g., Lemma III.1.2 in [37]).

**Theorem 5** *Let  $X \subset Y$  and let the operator  $M$  in (34) be of the form  $M = \mu I + T$ , where  $I$  is the identity in  $X$  and  $T$  is of class  $C^1$ . Furthermore, set  $L := D_x T(0)$ . Suppose that  $\mu_0 I + L$  is Fredholm of index 0 (Definition 2), for some  $\mu_0 \in U$ , and that  $-\mu_0$  is a simple eigenvalue for  $L$ , namely, the equation  $\mu_0 x + L(x) = 0$  has one and only one (nonzero) independent solution,  $x_1$ , while the equation  $\mu_0 x + L(x) = x_1$  has no solutions. Then,  $(0, \mu_0)$  is a bifurcation point for the equation  $M(x, \mu) = 0$ .*

**Bifurcation of Taylor–Couette Flow** A notable application of Theorem 5 is the *bifurcation of Taylor–Couette flow*. In this case the liquid fills the space between two coaxial, infinite cylinders,  $C_1$  and  $C_2$ , of radii  $R_1$  and  $R_2 > R_1$ , respectively.  $C_1$  rotates around the common axis,  $a$ , with constant angular velocity  $\omega$ , while  $C_2$  is at rest. Denote by  $(r, \theta, z)$  a system of cylindrical coordinates with  $z$  along  $a$  and oriented as  $\omega$  and let  $(e_r, e_\theta, e_z)$  the associated canonical base. The components of a vector  $w$  in such a base are denoted by  $w_r, w_\theta$  and  $w_z$ , respectively. If we introduce the non-dimensional quantities

$$u = v/(\omega r_1), \quad \tilde{x} = x/r_1, \quad p = p/(\rho \omega^2 r_1^2), \quad R = R_2/R_1,$$

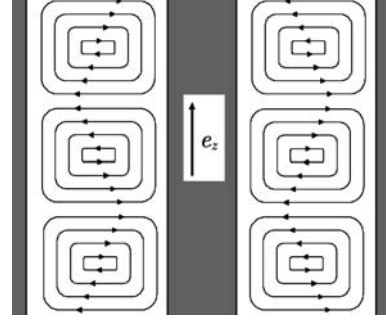
with  $\omega = |\omega|$ , we see at once that the following velocity and pressure fields

$$u_0 = (1 - R^2)^{-1} (r - R^2/r) e_\theta, \quad p_0 = |u_0|^2 \ln r + \text{const}, \quad (38)$$

solve (9), with  $f = 0$ , for all values of the Reynolds number  $\lambda := \rho \omega r_1^2 / \mu$ . Moreover,  $u_0$  satisfies the boundary conditions

$$u_0(r) = 0 \quad \text{at } r = 1, \quad u_0(r) = e_\theta \quad \text{at } r = R. \quad (39)$$

Experiments show that, if  $\lambda$  exceeds a critical value,  $\lambda_c$ , a flow with entirely different features than the flow (38) is observed. In fact, this new flow is dominated by large toroidal vortices, stacked one on top of the other, called *Taylor vortices*. They are periodic in the  $z$ -direction and axisymmetric (independent of  $\theta$ ); see Fig. 4. Therefore, we look for bifurcating solutions of the form  $u_0(r) + w(r, z)$ ,



**Navier–Stokes Equations: A Mathematical Analysis, Figure 4**  
Sketch of the streamlines of Taylor vortices, at the generic section  $\theta = \text{const}$

$p_0(r) + p(r, z)$ , satisfying (39), and where  $w$  and  $p$  are periodic in the  $z$ -direction with period  $\mathfrak{P}$ , and  $w$  satisfies the following parity conditions:

$$\begin{aligned} w_r(r, z) &= w_r(r, -z), & w_\theta(r, z) &= w_\theta(r, -z), \\ w_z(r, z) &= -w_z(r, -z). \end{aligned} \quad (40)$$

Moreover, the relevant region of flow becomes the *periodicity cell*  $\Omega := (1, R) \times (0, \mathfrak{P})$ . If we now introduce the *stream function*  $\psi$

$$\frac{\partial \psi}{\partial z} = w_r, \quad \frac{\partial(r\psi)}{\partial r} = -r w_z; \quad \psi(r, z) = -\psi(r, -z), \quad (41)$$

and the vector  $u := (\psi, w_\theta)$ , it can be shown (see § 72.7 in [99]) that  $u$  satisfies an equation of the type (36), namely,

$$u + \lambda (\bar{B}(v_0)(u) + \tilde{\mathcal{N}}(u)) = 0 \quad \text{in } \mathcal{H}(\Omega) \quad (42)$$

where the operators  $\bar{B}(v_0)$  and  $\tilde{\mathcal{N}}$  obey the same functional properties as  $B(v_0)$  and  $\mathcal{N}$ , and where  $\mathcal{H}(\Omega)$  is the Banach space of functions  $u := (\psi, w_\theta) \in C^{4,\alpha}(\Omega) \times C^{2,\alpha}(\Omega)$ ,  $\alpha \in (0, 1)$ , such that: (i)  $u(r, 0) = u(r, \mathfrak{P})$  for all  $r \in (1, R)$ , (ii)  $u$  satisfies the parity conditions in (40), (41), and (iii)  $\psi = \partial \psi / \partial r = w_\theta = 0$  at  $r = 1, R$ . Thus, in view of Theorem 5 and of the properties of the operators involved in (42), we will obtain that  $(0, \lambda_0)$  is a bifurcation point for (41) if the following two conditions are met

- (a)  $u + \lambda_0 \bar{B}(v_0)(u) = 0$   
has one and only one independent solution,  
 $u_1 \in \mathcal{H}(\Omega)$ ; (43)
- (b) the equation  $u + \lambda_0 \bar{B}(v_0)(u) = u_1$   
has no solution in  $\mathcal{H}(\Omega)$ .



It is in Lemma 72.14 in [99] that there exists a period  $\mathfrak{P}$  for which both conditions in (43) are satisfied. In addition, for all  $\lambda \in (0, \lambda_0)$  the equation  $\mathbf{u} + \lambda \tilde{\mathbf{B}}(\mathbf{v}_0)(\mathbf{u}) = \mathbf{0}$  has only the trivial solution in  $\mathcal{H}(\Omega)$ , which, by Lemma 4, implies that no bifurcation occurs for  $\lambda \in (0, \lambda_0)$ , which, in turn, means that the bifurcation is *supercritical*.

### Flow in Exterior Domains

One of the most significant questions in fluid mechanics is to determine the properties of the steady flow of a liquid past a body  $\mathcal{B}$  of simple symmetric shape (such as a sphere or cylinder), over the entire range of the Reynolds number  $\lambda := \rho UL/\mu \in (0, \infty)$ ; see, e.g., [§ 4.9] in [4]. Here  $L$  is a length scale representing the linear dimension of  $\mathcal{B}$ , while  $\mathbf{v}_\infty = -U\mathbf{e}_1$ ,  $U = \text{const} > 0$ , is the uniform velocity field of the liquid at large distances from  $\mathcal{B}$ . In the mathematical formulation of this problem, one assumes that the liquid fills the whole space,  $\Omega$ , outside the closure of the domain  $\mathcal{B} \subset \mathbb{R}^n$ ,  $n = 2, 3$ . Thus, by scaling  $\mathbf{v}$  by  $U$  and  $\mathbf{x}$  by  $L$ , from (9)–(10) we obtain that the steady flow past a body consists in solving the following non-dimensional exterior boundary-value problem

$$\left. \begin{aligned} -\Delta \mathbf{w} - \lambda \frac{\partial \mathbf{w}}{\partial x_1} + \lambda \mathbf{w} \cdot \nabla \mathbf{w} &= -\nabla p + \mathbf{f} \\ \text{div } \mathbf{w} &= 0 \end{aligned} \right\} \quad \text{in } \Omega \quad (44)$$

$$\mathbf{w}(\mathbf{x}) = \mathbf{e}_1, \quad \mathbf{x} \in \partial\Omega, \quad \lim_{|\mathbf{x}| \rightarrow \infty} \mathbf{w}(\mathbf{x}) = \mathbf{0},$$

where  $\mathbf{w} := \mathbf{v} + \mathbf{e}_1$  and where we assume, for simplicity, that  $\mathbf{v}_* \equiv \mathbf{0}$ . (However, all the stated results continue to hold, more generally, if  $\mathbf{v}_*$  belongs to a suitable trace space.)

Whereas in the three-dimensional case (e.g., *flow past a sphere*) the investigation of (44) is, to an extent, complete, in the two-dimensional case (e.g., *flow past a circular cylinder*) there are still fundamental unresolved issues. We shall, therefore, treat the two cases separately.

We need some preliminary considerations.

**Extension of the Boundary Data** If  $\Omega$  is smooth enough (locally Lipschitzian, for example), since  $\int_{\partial\Omega} \mathbf{e}_1 \cdot \mathbf{n} = 0$ , by what we observed in Sect. “Existence Results. Non-Homogeneous Boundary Conditions”, for any  $\lambda > 0$  we find  $\mathbf{V} = \mathbf{V}(\lambda, \mathbf{x})$ , with  $\text{div } \mathbf{V} = 0$  in  $\Omega$ , and satisfying (31) with  $\varepsilon = 1/(2\lambda)$ , say. Actually, as shown in Proposition 3.1 in [36], the extension  $\mathbf{V}$  can be chosen, in particular, to be of class  $C^\infty((0, \infty) \times \Omega)$  and such that the support of  $\mathbf{V}(\lambda, \cdot)$  is contained in a bounded set, independent of  $\lambda$ . The proof given in [36] is for  $\Omega \subset \mathbb{R}^3$ , but it can be easily extended to the case  $\Omega \subset \mathbb{R}^2$ .

**Variational Formulation and Weak Solutions** Setting  $\mathbf{u} = \mathbf{w} - \mathbf{V}$  in (44)<sub>1</sub>, dot-multiplying through both sides of this equation by  $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$ , integrating by parts and taking into account 12, we find

$$\begin{aligned} [\mathbf{u}, \boldsymbol{\varphi}] - \lambda \left( \frac{\partial \mathbf{u}}{\partial x_1}, \boldsymbol{\varphi} \right) - \lambda (\mathbf{u} \cdot \nabla \boldsymbol{\varphi}, \mathbf{u}) - \lambda \{ (\mathbf{u} \cdot \nabla \boldsymbol{\varphi}, \mathbf{V}) \\ + (\mathbf{V} \cdot \nabla \boldsymbol{\varphi}, \mathbf{u}) \} + (\mathbf{H}, \boldsymbol{\varphi}) = \langle \mathbf{f}, \boldsymbol{\varphi} \rangle, \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega). \end{aligned} \quad (45)$$

where

$$\mathbf{H} = \mathbf{H}(\mathbf{V}) := \Delta \mathbf{V} + \lambda \frac{\partial \mathbf{V}}{\partial x_1} - \lambda \mathbf{V} \cdot \nabla \mathbf{V}.$$

**Definition 5** A function  $\mathbf{w} \in D^{1,2}(\Omega)$  is called a weak solution to (44) if and only if  $\mathbf{w} = \mathbf{u} + \mathbf{V}$  where  $\mathbf{u} \in \mathcal{D}_0^{1,2}(\Omega)$  and  $\mathbf{u}$  satisfies (45).

### Regularity of Weak Solutions

- (A) *Local Regularity.* The proof of differentiability properties of weak solutions can be carried out in a way similar to the case of a bounded domain. In particular, if  $\mathbf{f}$  and  $\Omega$  are of class  $C^\infty$ , one can show that  $\mathbf{w} \in C^\infty(\bar{\Omega})$  and there exists a scalar field  $p \in C^\infty(\bar{\Omega})$  such that (44)<sub>1,2,3</sub> holds in the ordinary sense. For this and other intermediate regularity results, we refer to Theorem IX.1.1 in [32].
- (B) *Regularity at Infinity.* The study of the validity of (44)<sub>4</sub> and, more particularly, of the *asymptotic structure* of weak solutions, needs a much more involved treatment, and the results depend on whether the flow is three- or two-dimensional. In the *three-dimensional case*, if  $\mathbf{f}$  satisfies suitable summability assumptions outside a ball of large radius, by using a *local* representation of a weak solution, (namely, at all points of a ball of radius 1 and centered at  $\mathbf{x} \in \Omega$ ,  $\text{dist}(\mathbf{x}, \partial\Omega) > 1$ ) one can prove that  $\mathbf{w}$  and all its derivatives tend to zero at infinity uniformly pointwise and that a similar property holds for  $p$  (see Theorem X.6.1 in [32]). The starting point of this analysis is the crucial fact that

$$\begin{aligned} \text{if } \Omega \subset \mathbb{R}^3, \\ \mathcal{D}_0^{1,2}(\Omega) \text{ is continuously embedded in } L^6(\Omega), \end{aligned} \quad (46)$$

which implies that  $\mathbf{w}$  tends to zero at infinity in a suitable sense. However, the existence of the “wake” behind the body along with the *sharp* order of decay of  $\mathbf{w}$  and  $p$  requires a more complicated analysis based on

the *global* representation of a weak solution (namely, at all points outside a ball of sufficiently large radius) by means of the *Oseen fundamental tensor*, along with maximal regularity estimates for the solution of the linearized *Oseen problem*, this latter being obtained by suppressing the nonlinear term  $\mathbf{w} \cdot \nabla \mathbf{w}$  in (44) [§§ IX.6, IX.7 and IX.8] in [32]. In particular, one can prove the following result concerning the behavior of  $\mathbf{w}$ ; (see Theorems IX.7.1, IX.8.1 and Remark IX.8.1 in [32]).

**Theorem 6** *Let  $\Omega$  be a three-dimensional exterior domain of class  $C^2$ , and let  $\mathbf{w}$  be a weak solution to (44) corresponding to a given  $\mathbf{f} \in L^q(\Omega)$ , for all  $q \in (1, q_0]$ ,  $q_0 > 3$ . Then*

$$\mathbf{w} \in L^r(\Omega), \quad \text{if and only if } r > 2. \quad (47)$$

If, in addition,  $\mathbf{f}$  is of bounded support, then, denoting by  $\theta$  the angle made by a ray starting from the origin of coordinates (taken, without loss of generality, in  $\mathbb{R}^3 - \bar{\Omega}$ ) with the positively directed  $x_1$ -axis, we have

$$|\mathbf{w}(\mathbf{x})| \leq \frac{M}{|\mathbf{x}| [1 + |\mathbf{x}|(1 + \cos \theta)]}, \quad \mathbf{x} \in \Omega, \quad (48)$$

where  $M = M(\lambda, \Omega) > 0$ .

**Remark 6** Two significant consequences of Theorem 6 are: (i) the total kinetic energy of the flow past an obstacle ( $\equiv \frac{1}{2} \rho \|\mathbf{w}\|_2^2$ ) is infinite, see (47); and (ii) the asymptotic decay of  $\mathbf{w}$  is faster outside any semi-infinite cone with its axis coinciding with the negative  $x_1$ -axis (existence of the “wake”); see (48).

The study of the asymptotic properties of solutions in the two-dimensional case is deferred till Sect. “Two-Dimensional Flow. The Problem of Existence”.

**Three-Dimensional Flow. Existence of Solutions and Related Properties** As in the case of a bounded domain, we may use two different approaches to the study of existence of weak solutions.

**A. Finite-Dimensional Method** Assume  $\mathbf{f} \in \mathcal{D}_0^{-1,2}(\Omega)$ . With the same notation as in Subsect. “A. The Finite-Dimensional Method”, we look for “approximate solutions” to (45) of the form  $\mathbf{u}_N = \sum_{i=1}^N c_{iN} \boldsymbol{\psi}_i$ , where

$$\begin{aligned} & [\mathbf{u}_N, \boldsymbol{\psi}_k] - \lambda \left( \frac{\partial \mathbf{u}}{\partial N_{x_1}}, \boldsymbol{\psi}_k \right) - \lambda (\mathbf{u}_N \cdot \nabla \boldsymbol{\psi}_k, \mathbf{u}_N) \\ & - \lambda \{ (\mathbf{u}_N \cdot \nabla \boldsymbol{\psi}_k, \mathbf{V}) + (\mathbf{V} \cdot \nabla \boldsymbol{\psi}_k, \mathbf{u}_N) \} + (\mathbf{H}, \boldsymbol{\psi}_k) \\ & = \langle \mathbf{f}, \boldsymbol{\psi}_k \rangle, \quad k = 1, \dots, N. \end{aligned} \quad (49)$$

As in the case of a bounded domain, existence to the algebraic system (49), in the unknowns  $c_{iN}$ , will be achieved if we show the uniform estimate (28). By dot-multiplying through both sides of (49) by  $c_{kN}$ , by summing over  $k$  between 1 and  $N$  and by using Lemma 1 we find

$$|\mathbf{u}_N|_{1,2}^2 = \lambda (\mathbf{u}_N \cdot \nabla \mathbf{u}_N, \mathbf{V}) - (\mathbf{H}, \mathbf{u}_N) + \langle \mathbf{f}, \mathbf{u}_N \rangle. \quad (50)$$

From the properties of the extension of the boundary data, we have that  $\mathbf{V}$  satisfies (31) with  $\varepsilon = 1/(2\lambda)$  and that, moreover  $-(\mathbf{H}, \mathbf{u}_N) \leq C|\mathbf{u}_N|_{1,2}$ , for some  $C = C(\Omega, \lambda) > 0$ . Thus, from this and from (50) we deduce the uniform bound (28). This latter implies the existence of a subsequence  $\{\mathbf{u}_{N'}\}$  converging to some  $\mathbf{u} \in \mathcal{D}_0^{1,2}(\Omega)$  weakly. Moreover, by Rellich theorem,  $\mathbf{u}_{N'} \rightarrow \mathbf{u}$  in  $L^4(K)$ , where  $K := \Omega \cap \{|\mathbf{x}| > \rho\}$ , all sufficiently large and finite  $\rho > 0$  (see Theorem II.4.2 in [31]). Consequently, recalling that  $\boldsymbol{\psi}_k$  is of compact support in  $\Omega$ , we proceed as in the bounded domain case and take the limit  $N \equiv N' \rightarrow \infty$  in (49) to show that  $\mathbf{u}$  satisfies (45) with  $\boldsymbol{\varphi} \equiv \boldsymbol{\psi}_k$ . Successively, by taking into account that every  $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$  can be approximated in  $L^3(\Omega)$  by linear combinations of  $\boldsymbol{\psi}_k$  (see Lemma VII.2.1 in [31]) by (46) and by Lemma 1 we conclude that  $\mathbf{u}$  satisfies (45). Notice that, as in the case of flow in bounded domains, this method furnishes existence for any  $\lambda > 0$  and any  $\mathbf{f} \in \mathcal{D}_0^{-1,2}(\Omega)$ .

**B. The Function-Analytic Method** As in the case of flow in a bounded domain, Subsect. “The Function-Analytic Method”ion IV.1.(B), our objective is to rewrite (45) as a nonlinear operator equation in an appropriate Banach space, where the relevant operator satisfies the assumptions of Lemma 3. In this way, we may draw the same conclusions of Theorem 1 also in the case of a flow past an obstacle. However, *unlike the case of flow in a bounded domain*, the map

$$\boldsymbol{\varphi} \in \mathcal{D}(\Omega) \mapsto (\mathbf{u} \cdot \nabla \boldsymbol{\varphi}, \mathbf{u}) \in \mathbb{R}$$

can *not* be extended to a linear, bounded functional in  $\mathcal{D}_0^{1,2}(\Omega)$ , if  $\mathbf{u}$  merely belongs to  $\mathcal{D}_0^{1,2}(\Omega)$ . Analogous conclusion holds for the map  $\boldsymbol{\varphi} \in \mathcal{D}(\Omega) \mapsto (\partial \mathbf{u} / \partial x_1, \boldsymbol{\varphi}) \in \mathbb{R}$ . The reason is because, in an exterior domain, the Poincaré inequality (18) and, consequently, the embedding 21 are, in general, not true. It is thus necessary to consider the above functionals for  $\mathbf{u}$  in a space *strictly contained* in  $\mathcal{D}_0^{1,2}(\Omega)$ . Set

$$\|\mathbf{u}\| := \sup_{\boldsymbol{\varphi} \in \mathcal{D}(\Omega)} \frac{\left| \left( \frac{\partial \mathbf{u}}{\partial x_1}, \boldsymbol{\varphi} \right) \right|}{|\boldsymbol{\varphi}|_{1,2}}$$

and let

$$X = X(\Omega) := \left\{ \mathbf{u} \in \mathcal{D}_0^{1,2}(\Omega) : \|\mathbf{u}\| < \infty \right\}.$$

Clearly,  $X(\Omega)$  endowed with the norm  $|\cdot|_{1,2} + \|\cdot\|$  is a Banach space. Moreover,  $X(\Omega) \subset L^4(\Omega)$ , continuously; see Proposition 1.1 in [36]. We may thus conclude, by Riesz theorem, by Hölder inequality and by the properties of the extension  $\mathbf{V}$  that, for any  $\mathbf{u} \in X(\Omega)$ , there exist  $\mathbf{L}(\mathbf{u})$ ,  $\mathcal{M}(\mathbf{u})$ ,  $\mathcal{V}(\mathbf{u})$ , and  $\mathcal{H}(\lambda)$  in  $\mathcal{D}_0^{1,2}(\Omega)$  such that, for all  $\varphi \in \mathcal{D}(\Omega)$ ,

$$\begin{aligned} -\left(\frac{\partial \mathbf{u}}{\partial x_1}, \varphi\right) &= [\mathbf{L}(\mathbf{u}), \varphi]; \\ -(\mathbf{u} \cdot \nabla \varphi, \mathbf{u}) &= [\mathcal{M}(\mathbf{u}), \varphi]; \\ -(\mathbf{u} \cdot \nabla \varphi, \mathbf{V}) - (\mathbf{V} \cdot \nabla \varphi, \mathbf{u}) &= [\mathcal{V}(\mathbf{u}), \varphi]; \\ (\mathbf{H}, \varphi) &= [\mathcal{H}(\lambda), \varphi]. \end{aligned}$$

Consequently, we obtain that (45) is equivalent to the following equation

$$\mathbf{M}(\lambda, \mathbf{u}) = \mathbf{F} \quad \text{in } \mathcal{D}_0^{1,2}(\Omega), \quad (51)$$

where

$$\begin{aligned} \mathbf{M}: (\lambda, \mathbf{u}) &\in (0, \infty) \times X(\Omega) \\ \mapsto \mathbf{u} + \lambda (\mathbf{L}(\mathbf{u}) + \mathcal{V}(\mathbf{u}) + \mathcal{M}(\mathbf{u})) + \mathcal{H}(\lambda) &\in \mathcal{D}_0^{1,2}(\Omega). \end{aligned}$$

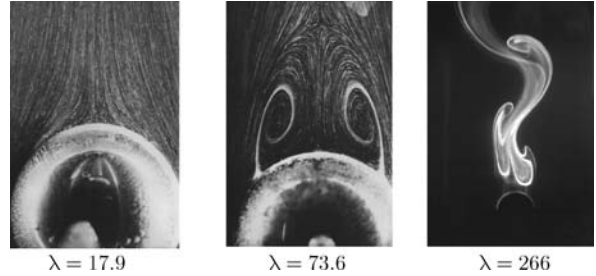
A detailed study of the properties of the operator  $\mathbf{M}$  is done in § 5 in [36], where, in particular, the following result is proved.

**Lemma 5** *The operator  $\mathbf{M}$  is of class  $C^\infty$ . Moreover, for each  $\lambda > 0$ ,  $\mathbf{M}(\lambda, \cdot) : X(\Omega) \mapsto \mathcal{D}_0^{1,2}(\Omega)$  is proper and Fredholm of index 0, and the two equations  $\mathbf{M}(\lambda, \mathbf{u}) = \mathcal{H}(\lambda)$  and  $D_u \mathbf{M}(\lambda, \mathbf{0})(\mathbf{w}) = \mathbf{0}$  only have the solutions  $\mathbf{u} = \mathbf{w} = \mathbf{0}$ .*

From this lemma and with the help of Lemma 3, we obtain the following result analogous to Theorem 1.

**Theorem 7** *For any  $\lambda > 0$  and  $\mathbf{F} \in \mathcal{D}_0^{1,2}(\Omega)$  the Eq. (51) has at least one solution  $\mathbf{u} \in X(\Omega)$ . Moreover, for each fixed  $\lambda > 0$ , there exists open and dense  $\mathcal{Q} = \mathcal{Q}(\lambda) \subset \mathcal{D}_0^{1,2}(\Omega)$  with the following properties: (i) For any  $\mathbf{F} \in \mathcal{Q}$  the number of solutions to (51),  $\mathfrak{n} = \mathfrak{n}(\mathbf{F}, \lambda)$  is finite and odd; (ii) the integer  $\mathfrak{n}$  is finite on each connected component of  $\mathcal{Q}$ .*

Finally, concerning the geometric structure of the set of pairs  $(\lambda, \mathbf{u}) \in (0, \infty) \times X(\Omega)$  satisfying (51) for a fixed  $\mathbf{F}$ , a result entirely similar to Theorem 2 continues to hold; see Theorem 6.2 in [36].



**Navier–Stokes Equations: A Mathematical Analysis, Figure 5**  
Visualization of a flow past a sphere at increasing Reynolds numbers  $\lambda$ ; after [91]

### Three-Dimensional Flow. Uniqueness and Steady Bifurcation

There is both experimental [91,97] and numerical [68,95] evidence that a steady flow past a sphere is unique (and stable) if the Reynolds number  $\lambda$  is sufficiently small. Moreover, experiments report that a closed recirculation zone first appears at  $\lambda$  around 20–25, and the flow stays steady and axisymmetric up to at least  $\lambda \simeq 130$ . This implies that the first bifurcation occurs through a steady motion. For higher values of  $\lambda$ , the wake behind the sphere becomes time-periodic, thus suggesting the occurrence of unsteady (Hopf) bifurcation; see Fig. 5.

In this section we shall collect the relevant results available for uniqueness and steady bifurcation of a flow past a three-dimensional obstacle.

**Uniqueness Results** Unlike the case of a flow in a bounded domain where uniqueness in the class of weak solutions is simply established (see Theorem 3), in the situation of a flow past an obstacle, the uniqueness proof requires the detailed study of the asymptotic behavior of a weak solution mentioned in Sect. “Flow in Exterior Domains”; see also Theorem 6. Precisely, we have the following result, for whose proof we refer to Theorem IX.5.3 in [32].

**Theorem 8** *Suppose  $\mathbf{f} \in L^{6/5}(\Omega) \cap L^{3/2}(\Omega)$ ,  $\lambda \in (0, \bar{\lambda}]$ , for some  $\bar{\lambda} > 0$ , and let  $\mathbf{w}$  be the corresponding weak solution. (Notice that, under the given assumptions,  $\mathbf{f} \in \mathcal{D}_0^{-1,2}(\Omega)$ .) There exists  $C = C(\Omega, \bar{\lambda}) > 0$  such that, if*

$$\|\mathbf{f}\|_{6/5} + \lambda < C,$$

*then  $\mathbf{w}$  is the only weak solution corresponding to  $\mathbf{f}$ .*

**Some Bifurcation Results** The rigorous study of steady bifurcation of a flow past a body is a very challenging mathematical problem. However, the function-analytic framework developed in the previous section allows

us to formulate sufficient conditions for steady bifurcation, that are *formally* analogous to those discussed in Sect. “Uniqueness and Steady Bifurcation” for the case of a flow in a bounded domain; see § VII in [36]. To this end, fix  $\mathbf{f} \in \mathcal{D}_0^{-1,2}(\Omega)$ , once and for all, and let  $\mathbf{u}_0 = \mathbf{u}_0(\lambda)$ ,  $\lambda$  in some open interval  $I \subseteq (0, \infty)$ , be a *given* curve in  $X(\Omega)$ , constituted by solutions to (51) corresponding to the prescribed  $\mathbf{f}$ . If  $\mathbf{u} + \mathbf{u}_0$ , is another solution, from (51) we easily obtain that  $\mathbf{u}$  satisfies the following equation in  $\mathcal{D}_0^{1,2}(\Omega)$

$$\mathbf{u} + \lambda (\mathbf{L}(\mathbf{u}) + \mathcal{B}(\mathbf{u}_0(\lambda))(\mathbf{u}) + \mathcal{M}(\mathbf{u})) = \mathbf{0}, \quad \mathbf{u} \in X(\Omega), \quad (52)$$

where  $\mathcal{B}(\mathbf{u}_0) := D_{\mathbf{u}}\mathcal{M}(\mathbf{u}_0)$ . In this setting, the branch  $\mathbf{u}_0(\lambda)$  becomes the solution  $\mathbf{u} \equiv \mathbf{0}$  and the bifurcation problem thus reduces to find a nonzero branch of solutions  $\mathbf{u} = \mathbf{u}(\lambda)$  to (52) in every neighborhood of some *bifurcation point*  $(\mathbf{0}, \lambda_0)$ ; see Definition 4. Define the map

$$\begin{aligned} F : (\lambda, \mathbf{u}) &\in (0, \infty) \times X(\Omega) \\ &\mapsto \mathbf{u} + \lambda (\mathbf{L}(\mathbf{u}) + \mathcal{B}(\mathbf{u}_0(\lambda))(\mathbf{u}) + \mathcal{M}(\mathbf{u})) \in \mathcal{D}_0^{1,2}(\Omega). \end{aligned} \quad (53)$$

In § VII in [36] the following result is shown.

**Lemma 6** *The map  $F$  is of class  $C^\infty$ . Moreover, the derivative*

$$D_{\mathbf{u}}F(\lambda, \mathbf{0})(\mathbf{w}) = \mathbf{w} + \lambda (\mathbf{L}(\mathbf{w}) + \mathcal{B}(\mathbf{u}_0(\lambda))(\mathbf{w})),$$

*is Fredholm of index 0.*

Therefore, from this lemma and from Lemma 4 we obtain that a *necessary condition* for  $(\mathbf{0}, \lambda_0)$  to be a bifurcation point to (52) is that the linear problem

$$\mathbf{w}_1 + \lambda_0 (\mathbf{L}(\mathbf{w}_1) + \mathcal{B}(\mathbf{u}_0(\lambda_0))(\mathbf{w}_1)) = \mathbf{0}, \quad \mathbf{w}_1 \in X(\Omega), \quad (54)$$

has a non-zero solution  $\mathbf{w}_1$ . Once this necessary condition is satisfied, one can formulate several sufficient conditions for the point  $(\mathbf{0}, \lambda_0)$  to be a bifurcation point. For a review of different criteria for global and local bifurcation for Fredholm maps of index 0, we refer to Sect. 6 in [33]. Here we wish to use the criterion of Theorem 5 to present a very simple (in principle) and familiar sufficient condition in the particular case when the given curve  $\mathbf{u}_0$  can be made (locally, in a neighborhood of  $\lambda_0$ ) independent of  $\lambda$ . This may depend on the particular non-dimensionalization of the Navier–Stokes equations and on the special form of the family of solutions  $\mathbf{u}_0$ . In fact, there are several interesting

problems formulated in exterior domains where this circumstance takes place, like, for example, the problem of steady bifurcation considered in the previous section and the one studied in Sect. 6 in [39]. Now, if  $\mathbf{u}_0$  does not depend on  $\lambda$ , from Theorem 5 and from (54) we immediately find that a sufficient condition in order that  $(\mathbf{0}, \lambda_0)$  be a bifurcation point is that the following problem

$$\mathbf{w} + \lambda_0 \mathcal{L}(\mathbf{w}) = \mathbf{w}_1, \quad \mathbf{w} \in X(\Omega), \quad (55)$$

with  $\mathcal{L} := \mathbf{L} + \mathcal{B}(\mathbf{u}_0)$  and  $\mathbf{w}_1$  solving (54), has no solution. In different words, *a sufficient condition for  $(\mathbf{0}, \lambda_0)$  to be a bifurcation point is that  $-1/\lambda_0$  is a simple eigenvalue of the operator  $\mathcal{L}$* . It is interesting to observe that this condition is *formally* the same as the one arising in steady bifurcation problems for flow in a *bounded* domain; see Sect. “Uniqueness and Steady Bifurcation”. However, while in this latter case  $\mathcal{L} (\equiv \mathcal{B}(\mathbf{v}_0))$  is *compact* and defined on the whole of  $\mathcal{D}_0^{1,2}(\Omega)$ , in the present situation  $\mathcal{L}$ , with domain  $D := X(\Omega) \subset \mathcal{D}_0^{1,2}(\Omega)$ , is an *unbounded* operator. As such, we can not even be sure that  $\mathcal{L}$  has real *simple* eigenvalues. Nevertheless, since, by Lemma 6,  $\mathbf{I} + \lambda \mathcal{L}$  is Fredholm of index 0 for all  $\lambda \in (0, \infty)$ , and since one can show (Lemma 7.1 in [36]) that  $\mathcal{L}$  is graph-closed, if  $\mathbf{u}_0 \in L^3(\Omega)$  (this latter condition is satisfied under suitable hypotheses on  $\mathbf{f}$ ; see Theorem 6) from well-known results of spectral theory (see Theorem XVII.2.1 in [46]), it follows that the set  $\Lambda$  of real eigenvalues of  $\mathcal{L}$  satisfies the following properties: (a)  $\Lambda$  is at most countable; (b)  $\Lambda$  is constituted by isolated points of finite algebraic and geometric multiplicities, and (c) points in  $\Lambda$  can only cluster at 0. Consequently, the bifurcation condition requiring the simplicity of  $-1/\lambda_0$  (namely, algebraic multiplicity 1) is perfectly meaningful.

## Two-Dimensional Flow. The Problem of Existence

The planar motion of a viscous liquid past a cylinder is among the oldest problems to have received a systematic mathematical treatment. Actually, in 1851, it was addressed by Sir George Stokes in his work on the motion of a pendulum in a viscous liquid [89]. In the wake of his successful study of the flow past a sphere in the limit of vanishing  $\lambda$  (*Stokes approximation*), Stokes looked for solutions to (44), with  $\mathbf{f} \equiv \mathbf{0}$  and  $\lambda = 0$ , in the case when  $\Omega$  is the exterior of a circle. However, to his surprise, he found that this linearized problem has *no* solution, and he concluded with the following (wrong) statement [89], p. 63,

“It appears that the supposition of steady motion is inadmissible”.

Such an observation constitutes what we currently call *Stokes Paradox*.



This is definitely a very intriguing starting point for the resolution of the boundary-value problem (44), in that it suggests that, if the problem has a solution, the nonlinear terms have to play a major role. In this regard, by using, for instance, the Galerkin method and proceeding exactly as in Subsection IV.2.1(A), we can prove the existence of a weak solution to (44), for any  $\lambda > 0$  and  $\mathbf{f} \in \mathcal{D}_0^{-1,2}(\Omega)$ . In addition, this solution is as smooth as allowed by the regularity of  $\Omega$  and  $\mathbf{f}$  (see Sect. “Flow in Exterior Domains”) and, in such a case, it satisfies (44)<sub>1,2,3</sub> in the ordinary sense. However, unlike the three-dimensional case [see Eq. (46)], the space  $\mathcal{D}_0^{1,2}(\Omega)$  is not embedded in any  $L^q$ -space and, therefore, we can not be sure that, even in an appropriate generalized sense, this solution vanishes at infinity, as requested by (44)<sub>4</sub>. Actually, if  $\Omega \subset \mathbb{R}^2$  there are functions in  $\mathcal{D}_0^{1,2}(\Omega)$  becoming unbounded at infinity. Take, for example,  $\mathbf{w} = (\ln |\mathbf{x}|)^\alpha \mathbf{e}_\theta$ ,  $\alpha \in (0, 1)$ , and  $\Omega$  the exterior of the unit circle. In this sense, we call these solutions *weak*, and not because of lack of local regularity (they are as smooth as allowed by the smoothness of  $\Omega$  and  $\mathbf{f}$ ). This problem was first pointed out by Leray (pp. 54–55 in [61]).

The above partial results leave open the worrisome possibility that a Stokes paradox could also hold for the fully nonlinear problem (45). If this chance turned out to be indeed true, it would cast serious doubts on the Navier–Stokes equations as a reliable fluid model, in that they would not be able to catch the physics of a very elementary phenomenon, easily reproduced experimentally.

The possibility of a nonlinear Stokes paradox was ruled out by Finn and Smith in a deep paper published in 1967 [20], where it is shown that if  $\mathbf{f} \equiv \mathbf{0}$  and if  $\Omega$  is sufficiently regular, then (45) has a solution, at least for “small” (but nonzero!)  $\lambda$ . The method used by these authors is based on the representation of solutions and careful estimates of the Green tensor of the Oseen problem. Another approach to existence for small  $\lambda$ , relying upon the  $L^q$  theory of the Oseen problem, was successively given by Galdi [30] where one can find the proof of the following result.

**Theorem 9** Let  $\Omega$  be of class  $C^2$  and let  $\mathbf{f} \in L^q(\Omega)$ , for some  $q \in (1, 6/5)$ . Then there exists  $\lambda_1 > 0$  and  $C = C(\Omega, q, \lambda_1) > 0$  such that if, for some  $\lambda \in (0, \lambda_1]$ ,

$$|\log \lambda|^{-1} + \lambda^{2(1/q-1)} \|\mathbf{f}\|_q < C,$$

problem (45) has at least one weak solution that, in addition, satisfies (45)<sub>4</sub> uniformly pointwise.

It can be further shown that the above solutions meet all the basic physical requirements. In particular, as in the

three-dimensional case (see Sect. “Flow in Exterior Domains”), they exhibit a “wake” in the region  $x_1 < 0$ . Moreover, the solutions are unique in a ball of a suitable Banach space, centered at the origin and of “small” radius. For the proof of these two statements, we refer to §§ X.4 and X.5 in [32].

Though significant, these results leave open several fundamental questions. The most important is, of course, that of whether problem (44) is solvable for all  $\lambda > 0$  and all  $\mathbf{f}$  in a suitable space. As we already noticed, this solvability would be secured if we can show that the weak solution does satisfy (44)<sub>4</sub>, even in a generalized sense. It should be emphasized that, since, as shown previously, there are functions in  $\mathcal{D}_0^{1,2}(\Omega)$  that become unbounded at infinity, the proof of this asymptotic property must be restricted to functions satisfying (44)<sub>1,2,3</sub>. This question has been taken up in a series of remarkable papers by Gilbarg and Weinberger [44], [45] and by Amick [2] in the case when  $\mathbf{f} \equiv \mathbf{0}$ . Some of their results lead to the following one due to Galdi; see Theorem 3.4 in [35].

**Theorem 10** Let  $\mathbf{w}$  be a weak solution to (44) with  $\mathbf{f} \equiv \mathbf{0}$ . Then, there exists  $\xi \in \mathbb{R}^2$  such that

$$\lim_{|\mathbf{x}| \rightarrow \infty} \mathbf{w}(\mathbf{x}) = \xi \quad \text{uniformly}.$$

**Open Question** It is not known if  $\xi = \mathbf{0}$ , and so it is not known if  $\mathbf{w}$  satisfies (44)<sub>4</sub>. Thus, the question of the solvability of (44) for arbitrary  $\lambda > 0$ , even when  $\mathbf{f} \equiv \mathbf{0}$ , remains open.

When  $\Omega$  is symmetric around the direction of  $\mathbf{e}_1$ , in [33] Galdi has suggested a different approach to the solvability of (44) (with  $\mathbf{f} \equiv \mathbf{0}$ ) for arbitrary large  $\lambda > 0$ . Let  $C$  be the class of vector fields,  $\mathbf{v} = (v_1, v_2)$ , and scalar fields  $\tau$  such that (i)  $v_1$  and  $\tau$  are even in  $x_2$  and  $v_2$  is odd in  $x_2$ , and (ii)  $|\mathbf{v}|_{1,2} < \infty$ . The following result holds.

**Theorem 11** Let  $\Omega$  be symmetric around the  $x_1$ -axis. Assume that the homogeneous problem:

$$\left. \begin{aligned} \Delta \mathbf{u} &= \mathbf{u} \cdot \nabla \mathbf{u} + \nabla \phi \\ \operatorname{div} \mathbf{u} &= 0 \end{aligned} \right\} \quad \text{in } \Omega \quad (56)$$

$$\mathbf{u}|_{\partial\Omega} = \mathbf{0}, \quad \lim_{|\mathbf{x}| \rightarrow \infty} \mathbf{u}(\mathbf{x}) = \mathbf{0} \quad \text{uniformly},$$

has only the zero solution,  $\mathbf{u} \equiv \mathbf{0}$ ,  $p = \text{const.}$ , in the class  $C$ . Then, there is a set  $M$  with the following properties:

- (i)  $M \subset (0, \infty)$ ;
- (ii)  $M \supset (0, c)$  for some  $c = c(\Omega) > 0$ ;
- (iii)  $M$  is unbounded;
- (iv) For any  $\lambda \in M$ , problem (44) has at least one solution in the class  $C$ .



**Open Question** The difficulty with the above theorem relies in establishing the validity of its hypothesis, namely, *whether or not (56) has only the zero solution in the class  $C$* . Moreover, supposing that the hypothesis holds true, the other fundamental problem is the study of the properties of the set  $M$ . For a detailed discussion of these issues, we refer to § 4.3 in [35].

### Mathematical Analysis of the Initial-Boundary Value Problem

Objective of this section is to present the main results and open questions regarding the unique solvability of the initial-boundary value problem (7)–(8), and significant related properties.

#### Preliminary Considerations

In order to present the basic problems for (7)–(8) and the related results, we shall, for the sake of simplicity, restrict ourselves to the case when  $f \equiv v_1 \equiv 0$ . In what follows, we shall denote by (7)–(8)<sub>hom</sub> this homogeneous problem. The first, fundamental problem that should be naturally set for (7)–(8)<sub>hom</sub> is the classical one of (global) well-posedness in the sense of Hadamard.

**Problem 1** Find a Banach space,  $X$ , such that for any initial data  $v_0$  in  $X$  there is a corresponding solution  $(v, p)$  to (7)–(8)<sub>hom</sub> satisfying the following conditions: (i) it exists for all  $T > 0$ , (ii) it is unique and (iii) it depends continuously on  $v_0$ .

In different words, the resolution of Problem 1 will ensure that the Navier–Stokes equations furnish, at all times, a *deterministic description* of the dynamics of the liquid, provided the initial data are given in a “sufficiently rich” class. It is immediately seen that the class  $X$  should meet some necessary requirements for Problem 1 to be solvable. For instance, if we take  $v_0$  *only bounded*, we find that problem (7)–(8), with  $\Omega = \mathbb{R}^n$  and  $f = 0$ , admits the following two distinct solutions

$$\begin{aligned} v_1(x, t) &= 0, & p_1(x, t) &= 0; \\ v_2(x, t) &= \sin t e_1, & p_2(x, t) &= -x_1 \cos t; \end{aligned}$$

corresponding to the same initial data  $v_0 = 0$ .

Furthermore, we observe that the resolution of Problem 1, does *not* exclude the possibility of the formation of a “singularity”, that is, the existence of points in the space-time region where the solution may become unboundedly large in certain norms. This possibility depends, of course, on the regularity of the functional class where well-posedness is established.

One is thus led to considering the next fundamental (and most popular) problem.

**Problem 2** Given an initial distribution of velocity  $v_0$ , no matter how smooth, with

$$\int_{\Omega} |v_0(x)|^2 < \infty, \quad (57)$$

determine a corresponding regular solution  $v(x, t)$ ,  $p(x, t)$  to (7)–(8)<sub>hom</sub> for all times  $t \in (0, T)$  and all  $T > 0$ .

By “regular” here we mean that  $v$  and  $p$  are both of class  $C^\infty$  in the *open cylinder*  $\Omega \times (0, T)$ , for all  $T > 0$ . When  $\Omega \equiv \mathbb{R}^3$ , Problem 1 is, basically, the third Millennium Prize Problem posted by the Clay Mathematical Institute in May 2000.

The requirement (57) on the initial data is meaningful from the physical point of view, in that it ensures that the kinetic energy of the liquid is initially finite. Moreover, it is also necessary from a mathematical viewpoint because, if we relax (57) to the requirement that the initial distribution of velocity is, for example, *only bounded*, then Problem 2 has a simple negative answer. In fact, the following pair

$$\begin{aligned} v(x, t) &= \frac{1}{\tau - t} e_1 & p(x, t) &= -\frac{x_1}{(\tau - t)^2}, \\ & & t \in [0, \tau), & \tau > 0 \end{aligned}$$

is a solution to (7)–(8)<sub>hom</sub> with  $f \equiv 0$  and  $\Omega = \mathbb{R}^3$ , that becomes singular at  $t = \tau$ , for any given positive  $\tau$ .

An alternative way of formulating Problem 2 in “more physical” terms is as follows.

**Problem 2'** Can a spontaneous singularity arise in a finite time in a viscous liquid that is initially in an arbitrarily smooth state?

Though, perhaps, the gut answer to this question could be in the negative, one can bring very simple examples of dissipative nonlinear evolution equations where spontaneous singularities do occur, if the initial data are sufficiently large. For instance, the initial-value problem  $v' + \sigma v = v^2$ ,  $v(0) = v_0$ ,  $\sigma > 0$ , has the explicit solution

$$v(t) = \frac{\sigma v_0}{v_0 - e^{\sigma t}(v_0 - \sigma)}$$

which shows that, if  $v_0 \leq \sigma$ , then  $v$  is smooth for all  $t > 0$ , while if  $v_0 > \sigma$ , then  $v$  becomes unbounded in a finite time:

$$v(t) \geq \frac{1}{\tau - t}, \quad t \in [0, \tau), \quad \tau := \frac{1}{\sigma} \log \left( \frac{v_0}{v_0 - \sigma} \right).$$

If the occurrence of singularities for problem (7)–(8)<sub>hom</sub> can not be at all excluded, one can still theorize that singularities are unstable and, therefore, “undetectable”. Another plausible explanation could be that singularity may appear in the Navier–Stokes equations due to the possible break-down of the continuum model at very small scales.

It turns out that, in the case of two-dimensional (2D) flow, both problems 1 and 2 are completely resolved, while they are both open in the three-dimensional (3D) case. The following section will be dedicated to these issues.

### On the Solvability of Problems 1 and 2

As in the steady-state case, a basic tool for the resolution of both Problems 1 and 2 is the accomplishment of “good” a priori estimates. By “good”, we mean that (i) they have to be *global*, namely, they should hold for all positive times, and (ii) they have to be valid in a sufficiently regular function class. These estimates can then be used along suitable “approximate solutions” which eventually will converge, by an appropriate limit procedure, to a solution to (7)–(8)<sub>hom</sub>. To date, “good” estimates for 3D flow are not known.

*Unless explicitly stated, throughout this section we assume that  $\Omega$  is a bounded, smooth (of class  $C^2$ , for example) domain of  $\mathbb{R}^n$ ,  $n = 2, 3$ .*

### Derivation of Some Fundamental A Priori Estimates

We recall that, for simplicity, we are assuming that the boundary data,  $\mathbf{v}_1$ , in (8) is vanishing. Thus, if we formally dot-multiply through both sides of (7)<sub>1</sub> (with  $\mathbf{f} \equiv \mathbf{0}$ ) by  $\mathbf{v}$ , integrate by parts over  $\Omega$  and take into account 12 and Lemma 1, we obtain the following equation

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{v}(t)\|_2^2 + \nu \|\mathbf{v}(t)\|_{1,2}^2 = 0. \quad (58)$$

The physical interpretation of (58) is straightforward. Actually, if we dot-multiply both sides of the identity  $\operatorname{div} \mathbf{D}(\mathbf{v}) = \Delta \mathbf{v}$  by  $\mathbf{v}$ , where  $\mathbf{D} = \mathbf{D}(\mathbf{v})$  is the stretching tensor (see 3), and integrate by parts over  $\Omega$ , we find that  $\|\mathbf{v}\|_{1,2} = \|\mathbf{D}(\mathbf{v})\|_2$ . Since, as we observed in Sect. “Derivation of the Navier–Stokes Equations and Preliminary Considerations”,  $\mathbf{D}$  takes into account the deformation of the parts of the liquid, Eq. (58) simply relates the rate of decreasing of the kinetic energy to the dissipation inside the liquid, due to the combined effect of viscosity and deformation. If we integrate (58) from  $s \geq 0$  to  $t \geq s$ , we obtain the so-called *energy equality*

$$\|\mathbf{v}(t)\|_2^2 + 2\nu \int_s^t \|\mathbf{v}(\rho)\|_{1,2}^2 d\rho = \|\mathbf{v}(s)\|_2^2 \quad 0 \leq s \leq t. \quad (59)$$

Notice that the nonlinear term  $\mathbf{v} \cdot \nabla \mathbf{v}$  does *not* give any contribution to Eq. (58) [and, consequently, to Eq. (59)], in virtue of the fact that  $(\mathbf{v} \cdot \nabla \mathbf{v}, \mathbf{v}) = 0$ ; see Lemma 1. By taking  $s = 0$  in (59), we find a *bound on the kinetic energy and on the total dissipation for all times  $t \geq 0$ , in terms of the initial data only*, provided the latter satisfy (57). In concise words, the energy equality (59) is a *global a priori* estimate. It should be emphasized that the energy equality is, basically, *the only known global a priori estimate for 3D flow*.

From (59) it follows, in particular,

$$\mathbf{v} \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; \mathcal{D}_0^{1,2}(\Omega)), \quad \text{all } T > 0. \quad (60)$$

A second estimate can be obtained by dot-multiplying through both sides of (7)<sub>1</sub> by  $P\Delta \mathbf{v}$  and by integrating by parts over  $\Omega$ . Taking into account that

$$\left( \frac{\partial \mathbf{v}}{\partial t}, P\Delta \mathbf{v} \right) = \left( \frac{\partial \mathbf{v}}{\partial t}, \Delta \mathbf{v} \right) = -\frac{1}{2} \frac{d}{dt} \|\mathbf{v}\|_{1,2}^2, \quad (\nabla p, P\Delta \mathbf{v}) = 0,$$

we deduce

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{v}\|_{1,2}^2 + \nu \|P\Delta \mathbf{v}\|_2^2 = (\mathbf{v} \cdot \nabla \mathbf{v}, P\Delta \mathbf{v}). \quad (61)$$

Since the right-hand side of this equation need not be zero, we get that, unlike (58), the nonlinear term *does* contribute to (61). In addition, since the sign of this contribution is basically unknown, in order to obtain useful estimates we have to increase it appropriately. To this end, we recall the validity of the following inequalities

$$\|\mathbf{u}\|_\infty \leq \begin{cases} c_1 \|\mathbf{u}\|_2^{\frac{1}{2}} \|P\Delta \mathbf{u}\|_2^{\frac{1}{2}} & \text{if } n = 2, \quad \text{for all } \mathbf{u} \in L_\sigma^2(\Omega), \\ & \text{with } P\Delta \mathbf{u} \in L^2(\Omega) \quad \text{and } \mathbf{u}|_{\partial\Omega} = \mathbf{0}, \\ c_2 \|\mathbf{u}\|_{1,2}^{\frac{1}{2}} \|P\Delta \mathbf{u}\|_2^{\frac{1}{2}} & \text{if } n = 3, \quad \text{for all } \mathbf{u} \in \mathcal{D}_0^{1,2}(\Omega), \\ & \text{with } P\Delta \mathbf{u} \in L^2(\Omega), \end{cases} \quad (62)$$

where  $c_i = c_i(\Omega) > 0$ ,  $i = 1, 2$ . These relations follow from the Sobolev embedding theorems along with Lemma 2. We shall sketch a proof in the case  $n = 3$ . By the property of the projection operator  $P$ ,  $\mathbf{u}$  satisfies the assumptions of Lemma 2 with  $\mathbf{g} := P\Delta \mathbf{u}$ , and, consequently, we have, on the one hand, that  $\mathbf{u} \in W^{2,2}(\Omega)$ , and, on the other hand,

$$\|\mathbf{u}\|_{2,2} \leq c \|P\Delta \mathbf{u}\|_2, \quad (63)$$

with  $c = c(\Omega) > 0$ . We now recall that there exists an extension operator  $E : \mathbf{u} \in W^{2,2}(\Omega) \mapsto E(\mathbf{u}) \in W^{2,2}(\mathbb{R}^3)$  such that

$$\|E(\mathbf{u})\|_{k,2} \leq C_k \|\mathbf{u}\|_{k,2}, \quad k = 0, 1, 2; \quad (64)$$

see Chap. VI, Theorem 5 in [88]. Next, take  $\boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^3)$ . From the identity  $\Delta(|\boldsymbol{\varphi}|^2) = 2(\boldsymbol{\varphi} \cdot \Delta \boldsymbol{\varphi} + |\nabla \boldsymbol{\varphi}|^2)$  we have the representation

$$|\boldsymbol{\varphi}(\mathbf{x})|^2 = -\frac{1}{2\pi} \int_{\mathbb{R}^3} \frac{\boldsymbol{\varphi}(\mathbf{y}) \cdot \Delta \boldsymbol{\varphi}(\mathbf{y}) + |\nabla \boldsymbol{\varphi}(\mathbf{y})|^2}{|\mathbf{x} - \mathbf{y}|} d\mathbf{y}. \quad (65)$$

Using Schwarz inequality on the right-hand side of (65) along with the classical Hardy inequality § II.5 in [31]:

$$\int_{\mathbb{R}^3} \frac{|\boldsymbol{\varphi}(\mathbf{y})|^2}{|\mathbf{x} - \mathbf{y}|^2} d\mathbf{y} \leq 4|\boldsymbol{\varphi}|_{1,2}^2,$$

we recover  $\|\boldsymbol{\varphi}\|_\infty \leq (2/\pi)^{1/2} |\boldsymbol{\varphi}|_{1,2}^{1/2} |\boldsymbol{\varphi}|_{2,2}^{1/2}$ . Since  $C_0^\infty(\mathbb{R}^3)$  is dense in  $W^{2,2}(\mathbb{R}^3)$ , from this latter inequality we deduce, in particular,

$$\|\mathbf{u}\|_\infty \leq (2/\pi)^{1/2} |E(\mathbf{u})|_{1,2}^{1/2} |E(\mathbf{u})|_{2,2}^{1/2},$$

which, in turn, by (64) and Poincaré's inequality (18) implies that

$$\|\mathbf{u}\|_\infty \leq c_5 |\mathbf{u}|_{1,2}^{1/2} \|\mathbf{u}\|_{2,2}^{1/2},$$

with  $c_5 = c_5(\Omega) > 0$ . Equation (62)<sub>2</sub> then follows from this latter inequality and from (63). (For the proof of (62) in more general domains, as well as in domains with less regularity, we refer to [9,66,98]. I am not aware of the validity of (62) in an arbitrary (smooth) domain.) We now employ (62) and 29 into (61) to obtain

$$\frac{1}{2} \frac{d}{dt} |\mathbf{v}|_{1,2}^2 + \frac{\nu}{2} \|P\Delta \mathbf{v}\|_2^2 \leq \begin{cases} c_3 \|\mathbf{v}\|_2^2 |\mathbf{v}|_{1,2}^4 & \text{if } n = 2, \\ c_4 |\mathbf{v}|_{1,2}^6 & \text{if } n = 3, \end{cases} \quad (66)$$

where  $c_i = c_i(\Omega, \nu) > 0$ ,  $i = 3, 4$ . Thus, observing that, from (59),  $\|\mathbf{v}(t)\|_2 \leq \|\mathbf{v}_0\|_2$ , if we assume, further, that  $\mathbf{v}_0 \in \mathcal{D}_0^{1,2}(\Omega)$ , from the previous differential inequality, we obtain the following uniform bound

$$|\mathbf{v}(t)|_{1,2} \leq M(\Omega, \nu, t, \|\mathbf{v}_0\|_{1,2}),$$

for all  $t \in [0, \tau)$ , and some  $\tau \geq 1/(K |\mathbf{v}_0|_{1,2}^\alpha)$ , (67)

where  $M$  is a continuous function in  $t$  and  $K = 8c_3$ ,  $\alpha = 2$  if  $n = 2$ , while  $K = 4c_4$ ,  $\alpha = 4$  if  $n = 3$ . Equation (67) provides the second a priori estimate. Notice that, unlike (59), we do not know if in (67) we can take  $t$  arbitrary large,

namely, we do not know if  $\tau = \infty$ . Integrating both sides of (66) from 0 to  $t < \tau$ , and taking into account (67) we find that

$$\int_0^t \|P\Delta \mathbf{v}(s)\|_2^2 ds \leq M_1(\Omega, \nu, t, \|\mathbf{v}_0\|_{1,2}),$$

for all  $t \in [0, \tau)$ , (68)

with  $M_1$  continuous in  $t$ . From (68), (67), (60), and 63 one can then show that

$$\mathbf{v} \in L^\infty(0, t; W_0^{1,2}(\Omega)) \cap L^2(0, t; W^{2,2}(\Omega)),$$

$t \in [0, \tau)$ . (69)

A third a priori estimate, on the time derivative of  $\mathbf{v}$ , can be formally obtained by dot-multiplying both sides of (7)<sub>1</sub> and by integrating by parts over  $\Omega$ . By using arguments similar to those leading to (61) we find

$$\frac{\nu}{2} \frac{d}{dt} |\mathbf{v}|_{1,2}^2 + \left\| \frac{\partial \mathbf{v}}{\partial t} \right\|_2^2 = - \left( \mathbf{v} \cdot \nabla \mathbf{v}, \frac{\partial \mathbf{v}}{\partial t} \right), \quad (70)$$

and so, employing Hölder inequality on the right-hand side of (62) along with 29, (67) and (68) we show the following estimate

$$\int_0^t \left\| \frac{\partial \mathbf{v}}{\partial s} \right\|_2^2 ds \leq M_2(\Omega, \nu, t, \|\mathbf{v}_0\|_{1,2}),$$

for all  $t \in [0, \tau)$ , (71)

with  $M_2$  continuous in  $t$ . This latter inequality implies that

$$\frac{\partial \mathbf{v}}{\partial t} \in L^2(0, t; L^2(\Omega)), \quad t \in [0, \tau). \quad (72)$$

**Existence, Uniqueness, Continuous Dependence, and Regularity Results** We shall now use estimates (59), (67), (68), (71) along a suitable approximate solution constructed by the finite-dimensional (Galerkin) method to show existence to (7)–(8)<sub>hom</sub> in an appropriate function class. We shall briefly sketch the argument. Similarly to the steady-state case, we look for an approximate solution to (7)–(8)<sub>hom</sub> of the form  $\mathbf{v}_N(\mathbf{x}, t) = \sum_{i=0}^N c_{iN}(t) \boldsymbol{\psi}_i$ , where  $\{\boldsymbol{\psi}_i\}$  is a base of  $L_\sigma^2(\Omega)$  constituted by the eigenvectors of the operator  $-P\Delta$ , namely,

$$-\nu \Delta \boldsymbol{\psi}_i = \lambda_i \boldsymbol{\psi}_i + \nabla \Phi_i, \quad \operatorname{div} \boldsymbol{\psi}_i = 0 \quad \text{in } \Omega,$$

$$\boldsymbol{\psi}_i|_{\partial\Omega} = \mathbf{0}, \quad i \in \mathbb{N}, \quad (73)$$

where  $\lambda_i$  are the corresponding eigenvalues. The coefficients  $c_{iN}(t)$  are requested to satisfy the following system

of ordinary differential equations

$$\left( \frac{\partial \mathbf{v}_N}{\partial t}, \boldsymbol{\psi}_k \right) + (\mathbf{v}_N \cdot \nabla \mathbf{v}_N, \boldsymbol{\psi}_k) = \nu (\Delta \mathbf{v}_N, \boldsymbol{\psi}_k),$$

$$k = 1, \dots, N, \quad (74)$$

with initial conditions  $c_{iN}(0) = (\mathbf{v}_0, \boldsymbol{\psi}_i)$ ,  $i = 1, \dots, N$ . Multiplying both sides of (74), in the order, by  $c_{kN}$ , by  $\lambda_k c_{kN}$  and by  $dc_{kN}/dt$  and summing over  $k$  from 1 to  $N$ , we at once obtain, with the help of (73) and of Lemma 1, that  $\mathbf{v}_N$  satisfies (59), (66), and (70). Consequently,  $\mathbf{v}_N$  satisfies the uniform (in  $N$ ) bounds (58) (evaluated at  $s = 0$ ) (67), (68), and (71). Employing these bounds together with, more or less, standard limiting procedures, we can show the existence of a field  $\mathbf{v}$  in the classes defined by (69) and (72) satisfying the relation

$$\left( \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} - \nu \Delta \mathbf{v}, \boldsymbol{\varphi} \right) = 0,$$

for all  $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$  and a.a.  $t \in [0, \tau)$ . (75)

Because of (69) and (72), the function involving  $\mathbf{v}$  in (75) belongs to  $L^2(\Omega)$ , a.e. in  $[0, \tau)$ , and therefore, in view of the orthogonal decomposition  $L^2(\Omega) = L_\sigma^2(\Omega) \oplus G(\Omega)$  and of the density of  $\mathcal{D}(\Omega)$  in  $L_\sigma^2(\Omega)$ , we find  $p \in L^2(0, t; W^{1,2}(\Omega))$ ,  $t \in [0, \tau)$ , such that  $(\mathbf{v}, p)$  satisfies (7)<sub>1</sub> for a.a.  $(\mathbf{x}, t) \in \Omega \times [0, \tau)$ . We thus find the following result, basically due to G. Prodi, to whose paper [72] we refer for all missing details; see also Chapter V.4 in [86].

**Theorem 12 (Existence)** *For every  $\mathbf{v}_0 \in \mathcal{D}_0^{1,2}(\Omega)$ , there exist  $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$  and  $p = p(\mathbf{x}, t)$  such that*

$$\left. \begin{aligned} \mathbf{v} &\in L^\infty(0, \tau; L^2(\Omega)) \cap L^2(0, \tau; \mathcal{D}_0^{1,2}(\Omega)), \\ \mathbf{v} &\in C([0, t]; \mathcal{D}_0^{1,2}(\Omega)) \\ &\cap L^2(0, t; W^{2,2}(\Omega)), \\ \frac{\partial \mathbf{v}}{\partial t} &\in L^2(0, t; L^2(\Omega)), \\ p &\in L^2(0, t; W^{1,2}(\Omega)), \end{aligned} \right\} \text{ for all } t \in [0, \tau), \quad (76)$$

with  $\tau$  given in (67), satisfying (7) for a.a.  $(\mathbf{x}, t) \in \Omega \times [0, \tau)$ , and (8)<sub>2</sub> (with  $\mathbf{v}_1 \equiv \mathbf{0}$ ) for a.a.  $(\mathbf{x}, t) \in \partial\Omega \times [0, \tau)$ . Moreover, the initial condition (8)<sub>1</sub> is attained in the following sense:

$$\lim_{t \rightarrow 0^+} \|\mathbf{v}(t) - \mathbf{v}_0\|_{1,2} = 0.$$

We also have.

**Theorem 13 (Uniqueness and Continuous Dependence on the Initial Data)** *Let  $\mathbf{v}_0$  be as in Theorem 12. Then the*

*corresponding solution is unique in the class (76). Moreover, it depends continuously on  $\mathbf{v}_0$  in the norm of  $L^2(\Omega)$ , in the time interval  $[0, \tau)$ .*

*Proof* Let  $(\mathbf{v}, p)$  and  $(\mathbf{v} + \mathbf{u}, p + p_1)$  be two solutions corresponding to data  $\mathbf{v}_0$  and  $\mathbf{v}_0 + \mathbf{u}_0$ , respectively. From (7)–(8)<sub>hom</sub> we then find

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} + \mathbf{v} \cdot \nabla \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{v} = -\nabla(p_1/\rho) + \nu \Delta \mathbf{u},$$

$$\operatorname{div} \mathbf{u} = 0, \quad \text{a.e. in } \Omega \times (0, \tau). \quad (77)$$

Employing the properties of the function  $\mathbf{v}$  and  $\mathbf{u}$ , it is not hard to show the following equation

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{u}\|_2^2 + \nu \|\mathbf{u}\|_{1,2}^2 = -(\mathbf{u} \cdot \nabla \mathbf{v}, \mathbf{u}), \quad (78)$$

that is formally obtained by dot-multiplying both sides of (77)<sub>1</sub> by  $\mathbf{u}$ , and by using (77)<sub>2</sub> and Lemma 1 along with the fact that  $\mathbf{u}$  has zero trace at  $\partial\Omega$ . By Hölder inequality and inequalities (14), (18), and (29), we find

$$\begin{aligned} |(\mathbf{u} \cdot \nabla \mathbf{v}, \mathbf{u})| &\leq \|\mathbf{u}\|_2 \|\mathbf{u}\|_4 \|\mathbf{v}\|_{1,4} \leq c_1 \|\mathbf{u}\|_2 \|\mathbf{u}\|_{1,2} \|\mathbf{v}\|_{2,2} \\ &\leq c_2 \|\mathbf{v}\|_{2,2}^2 \|\mathbf{u}\|_2^2 + \frac{\nu}{2} \|\mathbf{u}\|_{1,2}^2, \end{aligned}$$

where  $c_1 = c_1(\Omega) > 0$  and  $c_2 = c_2(\Omega, \nu) > 0$ . If we replace this inequality back into (78) we deduce

$$\frac{d}{dt} \|\mathbf{u}\|_2^2 \leq 2c_2 \|\mathbf{v}\|_{2,2}^2 \|\mathbf{u}\|_2^2,$$

and so, by Gronwall's lemma and by the fact that  $\int_0^t \|\mathbf{v}(s)\|_{2,2}^2 ds < \infty$ ,  $t \in [0, \tau)$  (see Theorem 12), we prove the desired result.  $\square$

Finally, we have the following result concerning the regularity of solutions determined in Theorem 12, for whose proof we refer to [38] and Theorem 5.2 in [34].

**Theorem 14 (Regularity)** *Let  $\Omega$  be a bounded domain of class  $C^\infty$ . Then, the solution  $(\mathbf{v}, p)$  constructed in Theorem 12 is of class  $C^\infty(\bar{\Omega} \times (0, \tau))$ .*

**Remark 7** The results of Theorems 12–14 can be extended to arbitrary domains of  $\mathbb{R}^n$ , provided their boundary is sufficiently smooth; see [34,50]. Moreover, the continuous dependence result in Theorem 13 can be proved in the stronger norm of  $\mathcal{D}_0^{1,2}(\Omega)$ .

**Times of Irregularity and Resolution of Problems 1 and 2 in 2D** Theorems 12–14 furnish a complete and positive answer to both Problems 1 and 2, provided we show that  $\tau = \infty$ . Our next task is to give necessary and sufficient

conditions for this latter situation to occur. To this end, we give the following.

**Definition 6** Let  $(\mathbf{v}, p)$  be a solution to (7)–(8)<sub>hom</sub> in the class (76). We say that  $\tau$  is a time of irregularity if and only if (i)  $\tau < \infty$ , and (ii)  $(\mathbf{v}, p)$  can not be continued, in the class (76), to an interval  $[0, \tau_1]$  with  $\tau_1 > \tau$ .

If  $\tau$  is a time of irregularity, we expect that some norms of the solution may become infinite at  $t = \tau$ , while being bounded for all  $t \in [0, \tau)$ . In order to show this rigorously, we premise a simple but useful result.

**Lemma 7** Assume that  $\mathbf{v}$  is a solution to (7)–(8)<sub>hom</sub> in the class (76) for some  $\tau > 0$ . Then,  $|\mathbf{v}(t)|_{1,2} < \infty$ , for all  $t \in [0, \tau)$ . Furthermore, for all  $q \in (n, \infty]$ , and all  $t \in [0, \tau)$

$$\int_0^t \|\mathbf{v}(s)\|_q^r ds < \infty, \quad \frac{2}{r} + \frac{n}{q} = 1.$$

*Proof* The proof of the first statement is obvious. Moreover, by the Sobolev embedding theorem (see Theorem at p. 125 in [70]) we find, for  $n = 2$ ,

$$\|\mathbf{v}\|_q \leq c_1 \|\mathbf{v}\|_2^{2/q} |\mathbf{v}|_{1,2}^{1-2/q}, \quad q \in (2, \infty) \quad (79)$$

and

$$\|\mathbf{v}\|_\infty \leq c_2 \|\mathbf{v}\|_{2,2},$$

with  $c_1 = c_1(q) > 0$ ,  $c_2 = c_2(\Omega) > 0$  while, if  $n = 3$ ,

$$\begin{aligned} \|\mathbf{v}\|_q &\leq c_3, \|\mathbf{v}\|_2^{(6-q)/2q} |\mathbf{v}|_{1,2}^{3(q-2)/2q}, \quad \text{if } q \in [2, 6], \\ \|\mathbf{v}\|_q^{2q/(q-3)} &\leq c_4 \|\mathbf{v}\|_{2,2}^{(q-6)/(q-3)} |\mathbf{v}|_{1,2}^{(6+q)/(q-3)}, \end{aligned} \quad (80)$$

if  $q \in (6, \infty]$ ,

where  $c_i = c_i(\Omega, q) > 0$ ,  $i = 3, 4$ . Since, by (76),

$$\sup_{t \in [0, \tau]} \|\mathbf{v}(t)\|_2 + |\mathbf{v}(t)|_{1,2} + \int_0^t \|\mathbf{v}(s)\|_{2,2}^2 ds < \infty,$$

$t \in [0, \tau)$ ,

the lemma follows by noting that  $(q-6)/(q-3) < 2$  for  $q > 3$ .  $\square$

We shall now furnish some characterization of the possible times of irregularity in terms of the behavior, around them, of the norms of the solution considered in Lemma 7.

**Lemma 8 (Criteria for the Existence of a Time of Irregularity)** Let  $(\mathbf{v}, p)$  be a solution to (7)–(8)<sub>hom</sub> in the

class (76) for some  $\tau \in (0, \infty]$ . Then, the following properties hold:

(i) If  $\tau$  is a time of irregularity, then

$$\lim_{t \rightarrow \tau^-} |\mathbf{v}(t)|_{1,2} = \infty. \quad (81)$$

Conversely, if  $\tau < \infty$  and (81) holds, then  $\tau$  is a time of irregularity. Moreover, if  $\tau$  is a time of irregularity, for all  $t \in (0, \tau)$  the following growth estimates hold

$$|\mathbf{v}(t)|_{1,2}^2 \geq \begin{cases} \frac{C}{\tau - t} & \text{if } n = 2, \\ \frac{C}{\sqrt{\tau - t}} & \text{if } n = 3, \end{cases} \quad (82)$$

with  $C = C(\Omega, \mathbf{v}) > 0$ .

(ii) If  $\tau$  is a time of irregularity, then, for all  $q \in (n, \infty]$ ,

$$\int_0^\tau \|\mathbf{v}(s)\|_q^r ds = \infty, \quad \frac{2}{r} + \frac{n}{q} = 1. \quad (83)$$

Conversely, if  $\tau < \infty$  and (83) holds for some  $q = \bar{q} \in (n, \infty]$ , then,  $\tau$  is a time of irregularity.

(iii) If  $n = 3$ , there exists  $K = K(\Omega, \mathbf{v}) > 0$  such that, if  $\|\mathbf{v}_0\|_2 |\mathbf{v}_0|_{1,2} < K$ , then  $\tau = \infty$ .

*Proof*

(i) Clearly, if  $\tau < \infty$  and (81) holds, then  $\tau$  is a time of irregularity. Conversely, suppose  $\tau$  is a time of irregularity and assume, by contradiction, that there exists a sequence  $\{t_k\}$  in  $[0, \tau)$  and  $M > 0$ , independent of  $k$ , such that

$$t_k \rightarrow \tau, \quad |\mathbf{v}(t_k)|_{1,2} \leq M.$$

Since  $\mathbf{v}(t_k) \in \mathcal{D}_0^{1,2}(\Omega)$ , by Theorem 12 we may construct a solution  $(\tilde{\mathbf{v}}, \tilde{p})$  with initial data  $\mathbf{v}(t_k)$ , in a time interval  $[t_k, t_k + \tau^*)$  where see (67)

$$\tau^* \geq A/|\mathbf{v}(t_k)|_{1,2}^\alpha \geq AM^\alpha \equiv \tau_*, \quad \alpha = 2(n-1),$$

and  $A$  depends only on  $\Omega$  and  $\mathbf{v}$ . By Theorem 12,  $\tilde{\mathbf{v}}$  belongs to the class (76) in the time interval  $[t_k, t_k + \tau_*]$ , with  $\tau_*$  independent of  $k$  and, by Theorem 13,  $\tilde{\mathbf{v}}$  must coincide with  $\mathbf{v}$  in the time interval  $[t_k, \tau)$ . We may now choose  $t_k \equiv \tau_0$  such that  $\tau_0 + \tau_* > \tau$ , contradicting the assumption that  $\tau$  is time of irregularity. We next show (82) when  $n = 3$ , the proof for  $n = 2$  being completely analogous. In-



tegrating (66) (with  $n = 3$ ) we find

$$\frac{1}{|\mathbf{v}(t)|_{1,2}^4} - \frac{1}{|\mathbf{v}(s)|_{1,2}^4} \leq c_4(s - t), \quad 0 < t < s < \tau.$$

Letting  $s \rightarrow \tau$  and recalling (81), we prove (82).

- (ii) Assume that  $\tau$  is a time of irregularity. Then, (82)<sub>2</sub> holds. Now, by the Sobolev embedding theorems, one can show that (see [proof of Lemma 5.4] in [34])

$$(\mathbf{v} \cdot \nabla \mathbf{v}, P\Delta \mathbf{v}) \leq C \|\mathbf{v}\|_q^{2q/(q-n)} |\mathbf{v}|_{1,2}^2 + \frac{\nu}{2} \|P\Delta \mathbf{v}\|_2^2, \\ \text{for all } q \in (n, \infty],$$

where  $C = C(\Omega, \nu) > 0$ . If we replace this relation into (61) and integrate the resulting differential inequality from 0 to  $t < \tau$ , we find

$$|\mathbf{v}(t)|_{1,2}^2 \leq |\mathbf{v}_0|_{1,2}^2 \exp\left\{2C \int_0^t \|\mathbf{v}(s)\|_q^r ds\right\}, \\ \text{for all } t \in [0, \tau]. \quad (84)$$

If condition (83) is not true for some  $q \in (n, \infty]$ , then (84) evaluated for that particular  $q$ , would contradict (82). Conversely, assume (83) holds for some  $q = \bar{q} \in (n, \infty]$ , but that, by contradiction, the solution of Theorem 12 can be extended to  $[0, \tau_1]$  with  $\tau_1 > \tau$ . Then, by Lemma 7 we would get the invalidity of condition (83) with  $q = \bar{q}$ , and the proof of (ii) is completed.

- (iii) By integrating the differential inequality (66)<sub>2</sub>, we find

$$|\mathbf{v}(t)|_{1,2}^2 \leq \frac{|\mathbf{v}_0|_{1,2}^2}{1 - 2c_4 |\mathbf{v}_0|_{1,2}^2 \int_0^t |\mathbf{v}(s)|_{1,2}^2 ds}, \quad t \in [0, \tau).$$

Thus, by (59) with  $s = 0$  in this latter inequality we find

$$|\mathbf{v}(t)|_{1,2}^2 \leq \frac{|\mathbf{v}_0|_{1,2}^2}{1 - c_5 |\mathbf{v}_0|_{1,2}^2 \|\mathbf{v}_0\|_2^2}, \quad t \in [0, \tau),$$

with  $c_5 = c_5(\Omega, \nu) > 0$ , which shows that (82)<sub>2</sub> can not occur if the initial data satisfy the imposed “smallness” restriction.

□

A fundamental consequence of Lemma 8 is that, in the case  $n = 2$ , a time of irregularity can not occur. In fact, for example, (82) is incompatible with the fact that  $\mathbf{v} \in L^2(0, \tau; \mathcal{D}_0^{1,2}(\Omega))$  (see (77)<sub>1</sub>). We thus have the following theorem, which answers positively both Problems 1 and 2 in the 2D case.

**Theorem 15 (Resolution of Problems 1 and 2 in 2D)** *Let  $\Omega \subset \mathbb{R}^2$ . Then, in Theorems 12–14 we can take  $\tau = \infty$ .*

The conclusion of Theorem 15 also follows from Lemma 8 (ii). In fact, in the case  $n = 2$ , by (79) we at once find that

$$L^{2q/(q-2)}(0, T; L^q(\Omega)) \subset L^\infty(0, T; L^2(\Omega)) \\ \cap L^2(0, T; \mathcal{D}_0^{1,2}(\Omega)), \text{ for all } T > 0 \text{ and all } q \in (2, \infty). \quad (85)$$

so that, from (76)<sub>1</sub>, we deduce

$$\int_0^\tau \|\mathbf{v}(t)\|_q^{2q/(2-q)} dt < \infty, \quad \text{for all } q \in (2, \infty).$$

which, by Lemma 8 (ii), excludes the occurrence of time of irregularity.

Unfortunately, from all we know, in the case  $n = 3$ , we can not draw the same conclusion. Actually, in such a case, (82)<sub>2</sub> and (77)<sub>1</sub> are no longer incompatible. Moreover, from Lemma 8 (ii) it follows that a sufficient condition for  $\tau$  not to be a time of irregularity is that

$$\int_0^\tau \|\mathbf{v}(t)\|_q^r dt < \infty, \\ \frac{2}{r} + \frac{3}{q} = 1, \quad \text{some } q \in (3, \infty]. \quad (86)$$

However, from (80)<sub>1</sub> and from (76), it is immediately verified that, in the case  $n = 3$ , the solutions constructed in Theorem 12 satisfy the following condition

$$\int_0^\tau \|\mathbf{v}(t)\|_q^r dt < \infty, \\ \frac{2}{r} + \frac{3}{q} = 1 + \frac{1}{2}, \quad \text{all } q \in [2, 6]. \quad (87)$$

Therefore, in view of Lemma 8 (iii), the best conclusion we can draw is that for 3D flow Problems 1 and 2 can be positively answered *if the size of the initial data is suitably restricted*.

**Remark 8** In the case  $n = 3$ , besides (86), one may furnish other sufficient conditions for the absence of a time of irregularity. We refer, among others, to the papers [6,8,15,18,56,57,69,80]. In particular, we would like to direct attention to the work [18,80], where the difficult borderline case  $q = n = 3$  in condition (86) is worked out by completely different methods than those used here.

**Open Question** In the case  $n = 3$ , it is not known whether or not condition (86) (or any of the other conditions referred to in Remark 8) holds along solutions of Theorem 12.

### Less Regular Solutions and Partial Regularity Results in 3D

As shown in the previous section, we do not know if, for 3D flow, the solutions of Theorem 12 exist in an arbitrarily large time interval, without restricting the magnitude of the initial data: they are *local solutions*. However, following the line of thought introduced by J. Leray [62], we may extend them to solutions defined for all times, for initial data of arbitrary magnitude, namely, to *global solutions*, but belonging to a functional class,  $C$ , a priori less regular than that given in (76) (*weak solutions*). Thus, if, besides existence, we could prove in  $C$  also uniqueness and continuous dependence, Problem 1 would receive a positive answer. Unfortunately, to date, the class where global solutions are proved to exist is, in principle, too large to secure the validity of these latter two properties and some extra assumptions are needed. Alternatively, in relation to Problem 2, one may investigate the “size” of the space-time regions where these generalized solutions may (possibly) become irregular. As a matter of fact, singularities, if they at all occur, have to be concentrated within “small” sets of space-time. Our objective in this section is to discuss the above issues and to present the main results.

For future purposes, we shall present some of these results also in space dimension  $n = 2$ , even though, as shown in Theorem 15, in this case, both Problems 1 and 2 are answered in the affirmative.

**Weak Solutions and Related Properties** We begin to introduce the corresponding definition of weak solutions in the sense of Leray–Hopf [54,62]. By formally dot-multiplying through both sides of (7) by  $\varphi \in \mathcal{D}(\Omega)$  and by integrating by parts over  $\Omega$ , with the help of 12 we find

$$\frac{d}{dt}(\mathbf{v}(t), \varphi) + \nu(\nabla \mathbf{v}(t), \nabla \varphi) + (\mathbf{v}(t) \cdot \nabla \mathbf{v}(t), \varphi) = 0, \quad \text{for all } \varphi \in \mathcal{D}(\Omega). \quad (88)$$

**Definition 7** Let  $\Omega \subset \mathbb{R}^n$ ,  $n = 2, 3$ . A field  $\mathbf{v} : \Omega \times (0, \infty) \mapsto \mathbb{R}^n$  is a weak solution to (7)–(8)<sub>hom</sub> if and only if: (i)  $\mathbf{v} \in L^\infty(0, T; L_\sigma^2(\Omega)) \cap L^2(0, T; \mathcal{D}_0^{1,2}(\Omega))$ , for all  $T > 0$ ; (ii)  $\mathbf{v}$  satisfies (88) for a.a.  $t \geq 0$ , and (iii)  $\lim_{t \rightarrow 0^+} (\mathbf{v}(t) - \mathbf{v}_0, \varphi) = 0$ , for all  $\varphi \in \mathcal{D}(\Omega)$ .

**A. Existence** The proof of existence of weak solutions is easily carried out, for example, by the finite-dimensional (Galerkin) method indicated in Subsection V.2.2. This time, however, along the “approximate” solutions  $\mathbf{v}_N$  to (74), we only use the estimate corresponding to the en-

ergy equality (59). We thus obtain

$$\|\mathbf{v}_N(t)\|_2^2 + 2\nu \int_s^t \|\mathbf{v}_N(\rho)\|_{1,2}^2 d\rho = \|\mathbf{v}_N(s)\|_2^2, \quad 0 \leq s \leq t. \quad (89)$$

As we already emphasized, the important feature of this estimate is that it holds for *all*  $t \geq 0$  and *all* data  $\mathbf{v}_0 \in L_\sigma^2(\Omega)$ . From (89) it follows, in particular, that the sequence  $\{\mathbf{v}_N\}$  is uniformly bounded in the class of functions specified in (i) of Definition 7. Using this fact together with classical weak and strong compactness arguments in (74), we can show the existence of at least one subsequence converging, in suitable topologies, to a weak solution  $\mathbf{v}$ . We then have the following result for whose complete proof we refer, e. g., to Theorem 3.1 in [34].

**Theorem 16** For any  $\mathbf{v}_0 \in L_\sigma^2(\Omega)$  there exists at least one weak solution to (7)–(8)<sub>hom</sub>. This solution verifies, in addition, the following properties.

(i) The energy inequality:

$$\|\mathbf{v}(t)\|_2^2 + 2\nu \int_s^t \|\mathbf{v}(\rho)\|_{1,2}^2 d\rho \leq \|\mathbf{v}(s)\|_2^2, \quad \text{for a.a. } s \in [0, \infty), 0 \text{ included, and all } t \geq s. \quad (90)$$

(ii)  $\lim_{t \rightarrow 0^+} \|\mathbf{v}(t) - \mathbf{v}_0\|_2 = 0$ .

**B. On the Energy Equality** In the case  $n = 3$ , weak solutions in Theorem 16 only satisfy the energy *inequality* (90) instead of the energy *equality* [see (59)]. (For the case  $n = 2$ , see Remark 9.) This is an undesired feature that is questionable from the physical viewpoint. As a matter of fact, for fixed  $s \geq 0$ , in time intervals  $[s, t]$  where (90) were to hold as a *strict* inequality, the kinetic energy would decrease by an amount which is not only due to the dissipation. From a strictly technical viewpoint, this happens because the convergence of (a subsequence of) the sequence  $\{\mathbf{v}_N\}$  to the weak solution  $\mathbf{v}$  can be proved only in the weak topology of  $L^2(0, T; \mathcal{D}_0^{1,2}(\Omega))$  and this only ensures that, as  $N \rightarrow \infty$ , the second term on the left-hand side of (89) tends to a quantity not less than the one given by the second term on the left-hand side of (90). One may think that this circumstance is due to the special method used for constructing the weak solutions. Actually, this is not the case because, in fact, we have the following.

**Open Question** If  $n = 3$ , it is not known if there are solutions satisfying (90) with the equality sign and corresponding to initial data  $\mathbf{v}_0 \in L_\sigma^2(\Omega)$  of unrestricted magnitude.

**Remark 9** A sufficient condition for a weak solution,  $\mathbf{v}$ , to satisfy the energy equality (59) is that  $\mathbf{v} \in L^4(0, t; L^4(\Omega))$ , for all  $t > 0$  (see Theorem 4.1 in [34]). Consequently, from (85) it follows that, if  $n = 2$ , weak solutions satisfy (59) for all  $t > 0$ . Moreover, from (80)<sub>1</sub>, we find that the solutions of Theorem 12, for  $n = 3$ , satisfy the energy equality (59), at least for all  $t \in [0, \tau]$ .

**Remark 10** For future purposes, we observe that the definition of weak solution and the results of Theorem 16 can be extended easily to the case when  $\mathbf{f} \neq \mathbf{0}$ . In fact, it is enough to change Definition 7 by requiring that  $\mathbf{v}$  satisfies the modification of (88) obtained by adding to its right-hand side the term  $(\mathbf{f}, \boldsymbol{\varphi})$ . Then, if  $\mathbf{f} \in L^2(0, T; \mathcal{D}_0^{-1,2}(\Omega))$ , for all  $T > 0$ , one can show the existence of a weak solution satisfying condition (ii) of Theorem 16 and the variant of (90) obtained by adding the term  $\int_0^t (\mathbf{f}, \mathbf{v}) \, ds$  on its right-hand side.

**C. Uniqueness and Continuous Dependence** The following result, due to Serrin (Theorem 6 in [83]) and Sather (Theorem 5.1 in [75]), is based on ideas of Leray [§ 32] in [62] and Prodi [71]. A detailed proof is given in Theorem 4.2 in [34].

**Theorem 17** Let  $\mathbf{v}, \mathbf{u}$  be two weak solutions corresponding to data  $\mathbf{v}_0$  and  $\mathbf{u}_0$ . Assume that  $\mathbf{u}$  satisfies the energy inequality (90) with  $s = 0$ , and that

$$\mathbf{v} \in L^r(0, T; L^q(\Omega)), \quad \text{for some } q \in (n, \infty] \\ \text{such that } \frac{2}{r} + \frac{n}{q} = 1. \quad (91)$$

Then,

$$\|\mathbf{v}(t) - \mathbf{u}(t)\|_2^2 \leq C \|\mathbf{v}_0 - \mathbf{u}_0\|_2^2 \exp \left\{ \int_0^t \|\mathbf{v}(\rho)\|_q^r \, d\rho \right\}, \\ \text{for all } t \in [0, T]$$

where  $C = C(\Omega, \nu) > 0$ . Thus, in particular, if  $\mathbf{v}_0 = \mathbf{u}_0$ , then  $\mathbf{v} = \mathbf{u}$  a.e. in  $\Omega \times [0, T]$ .

**Remark 11** If  $n = 2$ , from (85) and Remark 9 we find that every weak solution satisfies the assumptions of Theorem 17. Therefore, in such a case, every weak solution is unique in the class of weak solutions and depends continuously upon the data. Furthermore, if  $n = 3$ , the uniqueness result continues to hold if in condition (91), we take  $q = n = 3$ ; see [55, 87].

**Open Question** While the existence of weak solutions  $\mathbf{u}$  satisfying the hypothesis Theorem 17 is secured by The-

orem 16, in the case  $n = 3$  it is not known if there exist weak solutions having the property stated for  $\mathbf{v}$ . [In principle, as a consequence of Definition 7(i) and (80)<sub>1</sub>,  $\mathbf{v}$  only satisfies (87).] Consequently in the case  $n = 3$  uniqueness and continuous dependence in the class of weak solutions remains open, and so does the resolution of Problem 1.

**Remark 12** As a matter of fact, weak solutions possess more regularity than that implied by their very definition. Actually, if  $n = 2$ , they are indeed smooth (see Remark 13). If  $n = 3$ , by means of sharp estimates for solutions to the linear problem obtained from (7)–(8) by neglecting the nonlinear term  $\mathbf{v} \cdot \nabla \mathbf{v}$  (Stokes problem)—one can show that every corresponding weak solution satisfies the following additional properties (see Theorem 3.1 in [43], Theorem 3.4 in [87])

$$\frac{\partial \mathbf{v}}{\partial t} \in L^l(\delta, T; L^s(\Omega)), \quad \mathbf{v} \in L^l(\delta, T; W^{2,s}(\Omega)), \\ \text{for all } T > 0 \quad \text{and all } \delta \in (0, T),$$

where the numbers  $l, s$  obey the following conditions

$$\frac{2}{l} + \frac{3}{s} \geq 4, \quad l \in [1, 2], \quad s \in \left[1, \frac{3}{2}\right].$$

Moreover, there exists  $p \in L^l(\delta, T; W^{1,s}(\Omega)) \cap L^l(\delta, T; L^{3s/(3-s)}(\Omega))$  such that the pair  $(\mathbf{v}, p)$  satisfies (7) for a.a.  $(\mathbf{x}, t) \in \Omega \times (0, \infty)$ . If, in addition,  $\mathbf{v}_0$  lies in a sufficiently regular subspace of  $L_\sigma^2(\Omega)$ , we can take  $\delta = 0$ . However, the above properties are still not enough to ensure the validity of condition (91). Weak solutions enjoying further regularity properties are constructed in [14, 67] and Theorem 3.1 and Corollary 3.2 in [24].

#### D. Partial Regularity and “Suitable” Weak Solutions

A problem of fundamental importance is to investigate the set of space-time where weak solutions may possibly become irregular, and to give an estimate of how “big” this set can be.

To this end, we recall that, for a given  $S \subset \mathbb{R}^{d+1}$ ,  $d \in \mathbb{N} \cup \{0\}$ , and  $\kappa \in (0, \infty)$ , the  $\kappa$ -dimensional (spherical) Hausdorff measure  $\mathcal{H}^\kappa$  of  $S$  is defined as

$$\mathcal{H}^\kappa(S) = \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^\kappa(S),$$

where  $\mathcal{H}_\delta^\kappa(S) = \inf \sum_i r_i^\kappa$ , the infimum being taken over all at most countable coverings  $\{B_i\}$  of  $S$  of closed balls  $B_i \subset \mathbb{R}^d$  of radius  $r_i$  with  $r_i < \delta$ ; see, e.g., [19]. If  $d \in \mathbb{N}$ , the  $\kappa$ -dimensional parabolic Hausdorff measure  $\mathcal{P}^\kappa$  of  $S$  is defined as above, by replacing the ball  $B_i$  with a parabolic

cylinder of radius  $r_i$ :

$$Q_{r_i}(x, t) = \{(y, s) \in \mathbb{R}^d \times \mathbb{R} : |y - x| < r_i, |s - t| < r_i^2\}. \quad (92)$$

In general, it is  $\mathcal{H}^K(S) \leq C\mathcal{P}^K(S)$ ,  $C > 0$ ; see § 2.10.1 in [19] for details.

The following lemma is a direct consequence of the preceding definition.

**Lemma 9** *For any  $S \subset \mathbb{R}^{d+1}$ ,  $d \in \mathbb{N} \cup \{0\}$  (respectively,  $d \in \mathbb{N}$ ), we have  $\mathcal{H}^K(S) = 0$  (respectively,  $\mathcal{P}^K(S) = 0$ ) if and only if, for each  $\delta > 0$ ,  $S$  can be covered by closed balls  $\{B_i\}$  (respectively, parabolic cylinders  $\{Q_i\}$ ) of radii  $r_i$ ,  $i \in \mathbb{N}$ , such that  $\sum_{i=1}^{\infty} r_i^K < \delta$ .*

We begin to consider the collection of times where weak solutions are (possibly) not smooth and show that they constitute a very “small” region of  $(0, \infty)$ . Specifically, we have the following result, basically, due to Leray pp. 244–245 in [62] and completed by Scheffer [76].

**Theorem 18** *Let  $\Omega$  be a bounded domain of class  $C^\infty$ . Assume  $\mathbf{v}$  is a weak solution determined in Theorem 16. Then, there exists a union of disjoint and, at most, countable open time intervals  $\mathcal{T} = \mathcal{T}(\mathbf{v}) \subset (0, \infty)$  such that:*

- (i)  $\mathbf{v}$  is of class  $C^\infty$  in  $\bar{\Omega} \times \mathcal{T}$ ,
- (ii) There exists  $T^* \in (0, \infty)$  such that  $\mathcal{T} \supset (T^*, \infty)$ ;
- (iii) If  $\mathbf{v}_0 \in \mathcal{D}_0^{1,2}(\Omega)$  then  $\mathcal{T} \supset (0, T_1)$  for some  $T_1 > 0$ ;
- (iv) Let  $(s, \tau)$  be a generic bounded interval in  $\mathcal{T}(\mathbf{v})$  and suppose  $\mathbf{v} \notin C^\infty(\bar{\Omega} \times (s, \tau))$ ,  $\tau_1 > \tau$ . Then, both following conditions must hold

$$\begin{aligned} |\mathbf{v}(t)|_{1,2}^2 &\geq \frac{C}{(\tau - t)^{1/2}}, \quad t \in (s, \tau) \quad \text{and} \\ \lim_{t \rightarrow \tau^-} \int_s^t \|\mathbf{v}(s)\|_q^{2q/(q-3)} ds &= \infty, \quad \text{for all } q > 3, \end{aligned} \quad (93)$$

where  $C = C(\Omega, \mathbf{v}) > 0$ ;

- (v) The  $\frac{1}{2}$ -dimensional Hausdorff measure of  $\mathcal{I}(\mathbf{v}) := (0, \infty) - \mathcal{T}$  is zero;

*Proof* By (90) we may select  $T^* > 0$  with the following properties: (a)  $\|\mathbf{v}(T^*)\|_2 \|\mathbf{v}(T^*)\|_{1,2} < K$ , and (b) the energy inequality (90) holds with  $s = T^*$ , where  $K$  is the constant introduced in Lemma 8 (iii). Let us denote by  $\hat{\mathbf{v}}$  the solution of Theorem 12 corresponding to the data  $\mathbf{v}(T^*)$ . By Lemma 8 (iii),  $\hat{\mathbf{v}}$  exists for all times  $t \geq T^*$  and, by Theorem 14, it is of class  $C^\infty$  in  $\Omega \times (T^*, \infty)$ . By Lemma 7 and by Theorem 17 we must have  $\mathbf{v} \equiv \hat{\mathbf{v}}$  in  $\Omega \times (T^*, \infty)$ , and part (ii) is proved.

Next, denote by  $I$  the set of those  $t \in [0, T^*)$  such that (a)  $\|\mathbf{v}(t)\|_{1,2} < \infty$ , and (b) the energy inequality (90) holds with  $s \in I$ . Clearly,  $[0, T^*) - I$  is of zero Lebesgue measure. Moreover, for every  $t_0 \in I$  we can construct in  $(t_0, t_0 + T(t_0))$  a solution  $\hat{\mathbf{v}}$  assuming at  $t_0$  the initial data  $\mathbf{v}(t_0) \in \mathcal{D}_0^{1,2}(\Omega)$ ; see Theorem 12. From Theorem 14, Lemma 7, and Theorem 17, we know that  $\hat{\mathbf{v}}$  is of class  $C^\infty$  in  $\Omega \times (t_0, t_0 + T(t_0))$  and that it coincides with  $\mathbf{v}$ , since this latter satisfies the energy inequality with  $s = t_0$ . Furthermore, if  $\mathbf{v}_0 \in \mathcal{D}_0^{1,2}(\Omega)$ , then  $0 \in I$ . Properties (ii)–(iv) thus follow with  $\mathcal{T} \equiv \bigcup_{i \in \mathfrak{I}} (s_i, \tau_i) \cup (T^*, \infty)$ , where  $(s_i, \tau_i)$  are the connected components in  $I$ . Notice that

$$\begin{aligned} (s_i, \tau_i) &\subset [0, T^*], \quad \text{for all } i \in \mathfrak{I}; \\ (s_i, \tau_i) \cap (s_j, \tau_j) &= \emptyset, \quad i \neq j, \end{aligned} \quad (94)$$

and that, moreover, the (1-dimensional) Lebesgue measure of  $\mathcal{I} := \mathcal{T} - (0, \infty)$  is 0. Finally, property (iv) is an immediate consequence of Lemma 8 and Theorem 14. It remains to show (v). From (iv) and (90) we find

$$\begin{aligned} \sum_{i \in \mathfrak{I}} (\tau_i - s_i)^{1/2} &\leq 1/(2C) \\ \sum_{i \in \mathfrak{I}} \int_{\tau_i}^{s_i} \|\nabla \mathbf{v}(\tau)\|_2^2 d\tau &\leq \|\mathbf{v}_0\|_2^2/(4C). \end{aligned}$$

Thus, for every  $\delta > 0$  we can find a finite part  $\mathfrak{I}_\delta$  of  $\mathfrak{I}$  such that

$$\sum_{i \notin \mathfrak{I}_\delta} (\tau_i - s_i) < \delta, \quad \sum_{i \notin \mathfrak{I}_\delta} (\tau_i - s_i)^{1/2} < \delta. \quad (95)$$

By (94)<sub>1</sub>,  $\bigcup_{i \in \mathfrak{I}} (s_i, \tau_i) \subset [0, T^*]$  and so the set  $[0, T^*] - \bigcup_{i \in \mathfrak{I}_\delta} (s_i, \tau_i)$  consists of a finite number of disjoint closed intervals  $B_j$ ,  $j = 1, \dots, N$ . Clearly,

$$\bigcup_{j=1}^N B_j \supset \mathcal{I}(\mathbf{v}). \quad (96)$$

By (94)<sub>2</sub>, each interval  $(s_i, \tau_i)$ ,  $i \notin \mathfrak{I}_\delta$ , is included in one and only one  $B_j$ . Denote by  $\mathfrak{J}_j$  the set of all indices  $i$  satisfying  $B_j \supset (s_i, \tau_i)$ . We thus have

$$\mathfrak{I} = \mathfrak{I}_\delta \cup \left( \bigcup_{j=1}^N \mathfrak{J}_j \right), \quad (97)$$

$$B_j = \left( \bigcup_{i \in \mathfrak{J}_j} (s_i, \tau_i) \right) \cup (B_j \cap \mathcal{I}(\mathbf{v})).$$

Since  $\mathcal{I}$  has zero Lebesgue measure, from (97)<sub>2</sub> we have  $\text{diam } B_j = \sum_{i \in \mathfrak{J}_j} (\tau_i - s_i)$ . Thus, by (95) and (97)<sub>1</sub>,

$$\text{diam } B_j \leq \sum_{i \notin \mathfrak{I}_\delta} (\tau_i - s_i) < \delta \quad (98)$$

and, again by (95) and (97)<sub>1</sub>,

$$\begin{aligned} \sum_{j=1}^N (\text{diam } B_j)^{1/2} &\leq \sum_{j=1}^N \left( \sum_{i \in \mathfrak{J}_j} (\tau_i - s_i) \right)^{1/2} \\ &\leq \sum_{i \notin \mathfrak{J}_\delta} (\tau_i - s_i)^{1/2} < \delta. \end{aligned} \quad (99)$$

Therefore, property (v) follows from (96), (98), (99), and Lemma 9.  $\square$

**Remark 13** If  $\Omega$  is of class  $C^\infty$ , from 85, Remark 11 and from Theorem 18 (v) we at once obtain that, for  $n = 2$ , every weak solution is of class  $C^\infty(\bar{\Omega} \times (0, t))$ , for all  $t > 0$ .

We shall next analyze, in more detail, the set of points where weak solutions may possibly lose regularity. In view of Remark 13, we shall restrict ourselves to consider the case  $n=3$  only.

Let  $(s, \tau)$  be a bounded interval in  $\mathcal{T}(\mathbf{v})$  and assume that, at  $t = \tau$ , the weak solution  $\mathbf{v}$  becomes irregular. We may wish to estimate the spatial set  $\Sigma = \Sigma(\tau) \subseteq \Omega$  where  $\mathbf{v}(\tau)$  becomes irregular. By defining  $\Sigma$  as the set of  $\mathbf{x} \in \Omega$  where  $\mathbf{v}(\mathbf{x}, \tau)$  is not continuous, in the case  $\Omega = \mathbb{R}^3$ , Scheffer has shown that  $\mathcal{H}^1(\Sigma) < \infty$  [77].

More generally, one may wish to estimate the “size” of the region of space-time where points of irregularity (appropriately defined, see Definition 9 below) may occur. This study, initiated by Scheffer [77,78] and continued and deepened by Caffarelli, Kohn and Nirenberg [11], can be performed in a class of solutions called *suitable weak solutions* which, in principle, due to the lack of an adequate uniqueness theory, is more restricted than that of weak solutions.

**Definition 8** A pair  $(\mathbf{v}, p)$  is called a suitable weak solution to (7)–(8)<sub>hom</sub> if and only if: (i)  $\mathbf{v}$  satisfies Definition 7(i) and  $p \in L^{3/2}((0, T); L^{3/2}(\Omega))$ , for all  $T > 0$ ; (ii)  $(\mathbf{v}, p)$  satisfies

$$\begin{aligned} \frac{d}{dt}(\mathbf{v}(t), \boldsymbol{\psi}) + \nu(\nabla \mathbf{v}(t), \nabla \boldsymbol{\psi}) + (\mathbf{v}(t) \cdot \nabla \mathbf{v}(t), \boldsymbol{\psi}) \\ - (p(t), \text{div } \boldsymbol{\psi}) = 0, \quad \text{for all } \boldsymbol{\psi} \in C_0^\infty(\Omega), \end{aligned}$$

and (iii)  $(\mathbf{v}, p)$  obeys the following *localized energy inequality*

$$\begin{aligned} 2 \int_0^T \int_\Omega |\nabla \mathbf{v}|^2 \phi \, dx \, dt \\ \leq \int_0^T \int_\Omega \left\{ |\mathbf{v}|^2 \left( \frac{\partial \phi}{\partial t} + \Delta \phi \right) + (|\mathbf{v}|^2 + 2p) \mathbf{v} \cdot \nabla \phi \right\} dx \, dt, \end{aligned}$$

for all non-negative  $\phi \in C_0^\infty(\Omega \times (0, T))$ .

**Remark 14** By taking  $l = 3/2$ ,  $s = 9/8$  in Remark 12, it follows that every weak solution, corresponding to sufficiently regular initial data, matches requirements (i) and (ii) of Definition 8 (recall that  $\Omega$  is bounded). However, it is not known, to date, if it satisfies also condition (iii). Moreover, it is not clear if the finite-dimensional (Galerkin) method used for Theorem 16 is appropriate to construct solutions obeying such a condition (see, [47] for a partial answer). Nevertheless, by using different methods, one can show the existence of at least one weak solution satisfying the properties stated in Theorem 16, and which, in addition, is a suitable weak solution; (see [5], Theorem A.1 in [11], and Theorem 2.2 in [64]).

**Definition 9** A point  $P := (\mathbf{x}, t) \in \Omega_T := \Omega \times (0, T)$  is called regular for a suitable weak solution  $(\mathbf{v}, p)$ , if and only if there exists a neighborhood,  $I$ , of  $P$  such that  $\mathbf{v}$  is in  $L^\infty(I)$ . A point which is not regular will be called irregular.

**Remark 15** The above definition of a regular point is reinforced by a result of Serrin [82], from which we deduce that, in the neighborhood of every regular point, a suitable weak solution is, in fact, of class  $C^\infty$  in the space variables.

The next result is crucial in assessing the “size” of the set of possible irregular points in the space-time domain. For its proof, we refer to [64], Theorem 2.2 in [60], Proposition 2 in [11]. We recall that  $Q_r(\mathbf{x}, t)$  is defined in (92).

**Lemma 10** Let  $(\mathbf{v}, p)$  be a suitable weak solution and let  $(\mathbf{x}, t) \in \Omega_T$ . There exists  $K > 0$  such that, if

$$\limsup_{r \rightarrow 0} r^{-1} \int_{Q_r(\mathbf{x}, t)} |\nabla \mathbf{v}(\mathbf{y}, s)|^2 \, dy \, ds < K, \quad (100)$$

then  $(\mathbf{x}, t)$  is regular.

Now, let  $S = S(\mathbf{v}, p) \subseteq \Omega \times (0, T)$  be the set of possible irregular points for a suitable weak solution  $(\mathbf{v}, p)$ , and let  $V$  be a neighborhood of  $S$ . By Lemma 10 we then have that, for each  $\delta > 0$ , there is  $Q_r(\mathbf{x}, t) \subset V$  with  $r < \delta$ , such that

$$K^{-1} \int_{Q_r(\mathbf{x}, t)} |\nabla \mathbf{v}(\mathbf{y}, s)|^2 \, dy \, ds > r. \quad (101)$$

Let  $\mathcal{Q} = \{Q_r(\mathbf{x}, t)\}$  be the collection of all  $Q$  satisfying this property. Since  $\Omega \times (0, T)$  is bounded, from Lemma 6.1 in [11] we can find an at most countable, disjoint subfamily of  $\mathcal{Q}$ ,  $\{Q_{r_i}(\mathbf{x}_i, t_i)\}$ , such that  $S \subset \cup_{i \in \mathfrak{I}} Q_{5r_i}(\mathbf{x}_i, t_i)$ .



From (101) it follows, in particular, that

$$\begin{aligned} \sum_{i \in \mathfrak{S}} r_i &\leq K^{-1} \sum_{i \in \mathfrak{S}} \int_{Q_{r_i}(x_i, t_i)} |\nabla \mathbf{v}(\mathbf{y}, s)|^2 \, d\mathbf{y} \, ds \\ &\leq K^{-1} \int_V |\nabla \mathbf{v}|^2. \end{aligned} \quad (102)$$

Since  $\delta$  is arbitrary, 102 implies, on the one hand, that  $S$  is of zero Lebesgue measure and, on the other hand, that

$$\mathcal{P}^1(S) \leq \frac{5}{K} \int_V |\nabla \mathbf{v}|^2,$$

for every neighborhood  $V$  of  $S$ . Thus, by the absolute continuity of the Lebesgue integral, from this latter inequality and from Lemma 9 we have the following Theorem B in [11].

**Theorem 19** *Let  $(\mathbf{v}, p)$  be a suitable weak solution and let  $S = S(\mathbf{v}, p) \subseteq \Omega \times (0, T)$  be the corresponding set of possible irregular points. Then  $\mathcal{P}^1(S) = 0$ .*

**Remark 16** Let

$$\begin{aligned} \mathcal{Z} &= \mathcal{Z}(\mathbf{v}, p) \\ &:= \{t \in (0, T) : (\mathbf{x}, t) \in S(\mathbf{v}, p), \text{ for some } \mathbf{x} \in \Omega\}. \end{aligned}$$

Clearly,  $t \in \mathcal{Z}$  if and only if  $\mathbf{v}$  becomes essentially unbounded around  $(\mathbf{x}, t)$ , for some  $\mathbf{x} \in \Omega$ . Namely, for any  $M > 0$  there is a neighborhood of  $(\mathbf{x}, t)$ ,  $I_M$ , such that  $|\mathbf{v}(\mathbf{y}, s)| > M$  for a.a.  $(\mathbf{y}, s) \in I_M$ . From Theorem 18(ii) we deduce at once that  $\mathcal{Z}(\mathbf{v}, p) \subset \mathcal{I}(\mathbf{v})$ , where  $\mathcal{I}(\mathbf{v})$  is the set of all possible times of irregularity; see Definition 6. Thus, from Theorem 18(i) we find  $\mathcal{H}^{1/2}(\mathcal{Z}) = 0$ .

**Remark 17** There is a number of papers dedicated to the formulation of sufficient conditions for the absence of irregular points,  $(\mathbf{x}, t)$ , for a suitable weak solution, when  $\mathbf{x}$  is either an interior or a boundary point. In this latter case, the definition of the regular point as well as that of the suitable weak solution must be, of course, appropriately modified. Among others, we refer to [11, 48, 49, 64, 81].

### Long-Time Behavior and Existence of the Global Attractor

Suppose a viscous liquid moves in a fixed spatial bounded domain,  $\Omega$ , under the action of a given time-independent driving mechanism,  $m$ , and denote by  $\lambda > 0$  a non-dimensional parameter measuring the “magnitude” of  $m$ . To fix the ideas, we shall assume that the velocity field of the liquid,  $\mathbf{v}_1$ , vanishes at  $\partial\Omega$ , for each time  $t \geq 0$ . This assumption is made for the sake of simplicity. All

main results can be extended to the case  $\mathbf{v}_1 \not\equiv \mathbf{0}$ , provided  $\mathbf{v}_1(\mathbf{x}, t) \cdot \mathbf{n}|_{\partial\Omega} = 0$ , for all  $t \geq 0$ , where  $\mathbf{n}$  is the unit normal on  $\partial\Omega$ . We then take  $m$  to be a (non-conservative) time-independent body force,  $\mathbf{f} \in L^2(\Omega)$  with  $\lambda \sim \|\mathbf{f}\|_2$ . We shall denote by (7)–(8)<sub>Hom</sub> the initial-boundary value problem (7)–(8) with  $\mathbf{v}_1 \equiv \mathbf{0}$ . As we know from Theorem 3 and Remark 5, if  $\lambda$  is “sufficiently” small, less than  $\lambda_c$ , say, there exists one and only one (steady-state) solution,  $(\bar{\mathbf{v}}, \bar{p})$  to the boundary-value problem (9)–(10) (with  $\mathbf{v}_* \equiv \mathbf{0}$ ). Actually, it is easy to show that, with the above restriction on  $\lambda$ , every solution to the initial-boundary value problem (8)–(8)<sub>Hom</sub>, belonging to a sufficiently regular function class and corresponding to the given  $\mathbf{f}$  and to arbitrary  $\mathbf{v}_0 \in L^2_\sigma(\Omega)$ , decays exponentially fast in time to  $(\bar{\mathbf{v}}, \bar{p})$ . In fact, by setting  $\mathbf{u} := \mathbf{v} - \bar{\mathbf{v}}$ ,  $P := p - \bar{p}$ , from (7)–(10) we find

$$\begin{cases} \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} + \mathbf{u} \cdot \nabla \bar{\mathbf{v}} \\ \quad + \bar{\mathbf{v}} \cdot \nabla \mathbf{u} = -\nabla P + \nu \Delta \mathbf{u}, \\ \operatorname{div} \mathbf{u} = 0 \end{cases} \quad \text{in } \Omega \times (0, T);$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{v}_0(\mathbf{x}) - \bar{\mathbf{v}}(\mathbf{x}), \quad \mathbf{x} \in \Omega; \quad \mathbf{u} = \mathbf{0} \text{ at } \partial\Omega \times (0, T). \quad (103)$$

If we formally dot-multiply through both sides of (103)<sub>1</sub> by  $\mathbf{u}$ , integrate by parts over  $\Omega$  and take into account 12 and Lemma 1, we obtain

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{u}\|_2^2 + \nu \|\mathbf{u}\|_{1,2}^2 = -(\mathbf{u} \cdot \nabla \bar{\mathbf{v}}, \mathbf{u}). \quad (104)$$

From Lemma 1, (21), and (18) we find  $|(\mathbf{u} \cdot \nabla \bar{\mathbf{v}}, \mathbf{u})| \leq c_1 \|\mathbf{u}\|_{1,2}^2 \|\bar{\mathbf{v}}\|_{1,2}$ , with  $c_1 = c_1(\Omega) > 0$ . Thus, by Remark 5, it follows that

$$|(\mathbf{u} \cdot \nabla \bar{\mathbf{v}}, \mathbf{u})| \leq \frac{c_2}{\nu} \|\mathbf{f}\|_2 \|\mathbf{u}\|_{1,2}^2,$$

with  $c_2 = c_2(\Omega) > 0$ , which, in turn, once replaced in (104), furnishes

$$\frac{1}{2} \frac{d}{dt} \|\mathbf{u}\|_2^2 + \left( \nu - \frac{c_2}{\nu} \|\mathbf{f}\|_2 \right) \|\mathbf{u}\|_{1,2}^2 \leq 0.$$

Therefore, if  $\gamma := \nu - (c_2/\nu) \|\mathbf{f}\|_2 > 0$ , from (18) and from the latter displayed equation we deduce

$$\|\mathbf{u}(t)\|_2^2 \leq \|\mathbf{u}(0)\|_2^2 e^{-2(\gamma/C_P)t}, \quad (105)$$

which gives the desired result. Estimate (105) can be read, in “physical terms” in the following way: after a certain amount of time (depending on how close  $\gamma$  is to 0, namely, on how close  $\lambda$  is to  $\lambda_c$ ), the transient motion will die out exponentially fast and the “true” dynamics of the fluid will

be described by the unique steady-state flow corresponding to the given force  $\mathbf{f}$ . From a mathematical point of view, the steady-state  $(\bar{\mathbf{v}}, \bar{p})$  is, in a suitable function space, a one-point set which is invariant under the flow. Now, let us increase  $\lambda$  higher and higher beyond  $\lambda_c$ . Following Eberhard Hopf [53], we expect that, after a while, the transient motion will yet die out, and that the generic flow will approach a certain manifold,  $\mathfrak{M} = \mathfrak{M}(\lambda)$  which need not reduce to a single point. (There are, however, explicit examples where  $\mathfrak{M}(\lambda)$  remains a single point for any  $\lambda > 0$ ; see [65].) Actually, in principle, the structure of  $\mathfrak{M}$  can be very involved. Nevertheless, we envisage that  $\mathfrak{M}$  is still invariant under the flow, and that it is  $\mathfrak{M}$  where, eventually, the “true” dynamics of the liquid will take place. For obvious reasons, the manifold  $\mathfrak{M}$  is called the *global attractor*.

The existence of a global attractor and the study of the dynamics of the liquid on it, could be of the utmost importance in the effort of formulating a mathematical theory of *turbulence*. Actually, as is well known, if the magnitude of the driving force becomes sufficiently large, the corresponding flow becomes chaotic and the velocity and pressure of the liquid exhibit large and completely random variation in space and time. According to the ideas proposed by Smale [85] and by Ruelle and Takens [74], this chaotic behavior could be explained by the existence of a very complicated global attractor, where, as mentioned before, the ultimate dynamics of the liquid occurs.

**Existence of the Global Attractor for Two-Dimensional Flow, and Related Properties** Throughout this section we shall consider two-dimensional flow, so that, in particular,  $\Omega \subset \mathbb{R}^2$ .

Let  $\mathbf{f} \in L^2(\Omega)$  and  $\nu > 0$  be given. Consider the one-parameter family of operators

$$S_t : \mathbf{a} \in L^2_\sigma(\Omega) \mapsto S_t(\mathbf{a}) := \mathbf{v}(t) \in L^2_\sigma(\Omega), \quad t \in [0, \infty)$$

where  $\mathbf{v}(t)$  is, at each  $t$ , the weak solution to (7)–(8)<sub>Hom</sub> corresponding to the initial data  $\mathbf{a}$ ; see Remark 10. From Remark 9 and Remark 11 we deduce that the family  $\{S_t\}_{t \geq 0}$  defines a (strongly) *continuous semi-group* in  $L^2_\sigma(\Omega)$ , namely (i)  $S_{t_1}S_{t_2}(\mathbf{a}) = S_{t_1+t_2}(\mathbf{a})$  for all  $\mathbf{a} \in L^2_\sigma(\Omega)$  and all  $t_1, t_2 \in [0, \infty)$ ; (ii)  $S_0(\mathbf{a}) = \mathbf{a}$ , and (iii) the map  $t \mapsto S_t(\mathbf{a})$  is continuous for all  $\mathbf{a} \in L^2_\sigma(\Omega)$ .

**Definition 10** For any given  $\mathbf{f} \in L^2(\Omega)$ , the corresponding pair  $\{L^2_\sigma(\Omega), S_t\}$  is called *semi-flow* associated to (7)–(8)<sub>Hom</sub>.

Our objective is to study the asymptotic properties (as  $t \rightarrow \infty$ ) of the semi-flow  $\{L^2_\sigma(\Omega), S_t\}$ . To this end, we need to recall some basic facts.

Given  $\mathcal{A}_1, \mathcal{A}_2 \subset L^2_\sigma(\Omega)$ , we set

$$\delta(\mathcal{A}_1, \mathcal{A}_2) = \sup_{\mathbf{u}_1 \in \mathcal{A}_1} \inf_{\mathbf{u}_2 \in \mathcal{A}_2} \|\mathbf{u}_1 - \mathbf{u}_2\|_2.$$

Notice that  $\delta(\mathcal{A}_1, \mathcal{A}_2) = 0 \Rightarrow \mathcal{A}_1 \subseteq \bar{\mathcal{A}}_2$ . Moreover, we denote by  $\mathfrak{A}$  the class of all bounded subset of  $L^2_\sigma(\Omega)$ .

**Definition 11**  $\mathcal{B} \subset L^2_\sigma(\Omega)$  is called: (i) *absorbing* iff for any  $\mathcal{A} \in \mathfrak{A}$  there is  $t_0 = t_0(\mathcal{A}) \geq 0$  such that  $S_t(\mathcal{A}) \subseteq \mathcal{B}$ , for all  $t \geq t_0$ ; (ii) *attracting* iff  $\lim_{t \rightarrow \infty} \delta(S_t(\mathcal{A}), \mathcal{B}) = 0$ , for all  $\mathcal{A} \in \mathfrak{A}$ ; (iii) *invariant* iff  $S_t(\mathcal{B}) = \mathcal{B}$ , for all  $t \geq 0$ ; (iv) *maximal invariant* iff it is invariant and contains every invariant set in  $\mathfrak{A}$ ; (v) *global attractor* iff it is compact, attracting and maximal invariant.

Clearly, if a global attractor exists, it is unique. Furthermore, roughly speaking, its existence is secured whenever the semiflow admits a bounded absorbing set on which the semiflow becomes, eventually, relatively compact; (see Theorem 1.1 in [94]). The following result holds.

**Theorem 20** For any  $\mathbf{f} \in L^2_\sigma(\Omega)$  and  $\nu > 0$ , the corresponding semi-flow  $\{L^2_\sigma(\Omega), S_t\}$  admits a global attractor  $\mathfrak{M} = \mathfrak{M}(\mathbf{f}, \nu)$  which is also connected.

*Proof* In view of Theorem 1.1 in [94] and of Rellich compactness theorem (Theorem II.4.2 in [31]), it suffices to show the following two properties: (a) existence of a bounded absorbing set, and (b) given  $M > 0$ , there is  $t_0 = t_0(M, \mathbf{f}, \nu) > 0$  such that  $\|\mathbf{a}\|_2 \leq M$  implies  $|S_t(\mathbf{a})|_{1,2} \leq C$ , for all  $t \geq t_0$  and for some  $C > 0$  independent of  $t$ . The starting point is the analog of (58) and (61) which, this time, take the form

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|\mathbf{v}(t)\|_2^2 + \nu |\mathbf{v}(t)|_{1,2}^2 &= (\mathbf{f}, \mathbf{v}), \\ \frac{1}{2} \frac{d}{dt} |\mathbf{v}(t)|_{1,2}^2 + \nu \|P\mathbf{v}(t)\|_2^2 &= (\mathbf{v} \cdot \nabla \mathbf{v}, P\Delta \mathbf{v}) - (\mathbf{f}, P\Delta \mathbf{v}) \end{aligned} \quad (106)$$

By using in these equations the Schwarz inequality, and inequalities (18), 29, (62)<sub>1</sub> and (63), we deduce

$$\begin{aligned} \frac{d}{dt} \|\mathbf{v}(t)\|_2^2 + (\nu/C_P) \|\mathbf{v}(t)\|_2^2 &\leq F, \\ \frac{d}{dt} |\mathbf{v}(t)|_{1,2}^2 - g(t) |\mathbf{v}(t)|_{1,2}^2 &\leq F, \end{aligned} \quad (107)$$

where  $F := \|\mathbf{f}\|_2^2/\nu$ ,  $g(t) := c_1 \|\mathbf{v}(t)\|_2^2 |\mathbf{v}(t)|_{1,2}^2$ , and  $c_1 = c_1(\Omega) > 0$ . By integrating (107)<sub>1</sub>, we find

$$\|S_t(\mathbf{a})\|_2^2 := \|\mathbf{v}(t)\|_2^2 \leq \|\mathbf{a}\|_2^2 e^{-\nu t/C_P} + (1 - e^{-\nu t/C_P}) F/C_P. \quad (108)$$

Thus, setting

$$\mathfrak{B} := \{\varphi \in L^2_\sigma(\Omega) : \|\varphi\|_2 \leq (2F/C_P)^{1/2} \equiv \rho\}, \quad (109)$$

from (107) we deduce that, whenever  $\|\mathbf{a}\|_2 < M$ , there exists  $t_1 = t_1(M, F, \nu) > 0$  such that  $S_t(\mathbf{a}) \in \mathfrak{B}$ , which shows that  $\mathfrak{B}$  is absorbing. Next, again from Schwarz inequality and from (98)<sub>1</sub>, we obtain

$$\begin{aligned} \int_t^{t+1} |\mathbf{v}(s)|_{1,2}^2 ds &\leq \frac{1}{\nu} \|\mathbf{f}\|_2 \\ \int_t^{t+1} \|\mathbf{v}(s)\|_2 ds &\leq \left( \frac{\rho^4 M^2 C_P}{2\nu^2} \right)^{1/2} \equiv \rho_1 \quad \text{for all } t \geq t_1, \end{aligned} \quad (110)$$

which, in particular, implies

$$|\mathbf{v}(\bar{t})|_{1,2}^2 \leq \rho_1^2, \quad \text{for some } \bar{t} \in (t, t+1), \quad \text{all } t \geq t_1. \quad (111)$$

We next integrate (107)<sub>2</sub> from  $\bar{t}$  to  $t+1$  and use (110), to get

$$\begin{aligned} |\mathbf{v}(t+1)|_{1,2}^2 &\leq (\rho_1^2 + F) e^{\int_{\bar{t}}^{t+1} g(\zeta) d\zeta} \\ &\leq (\rho_1^2 + F) e^{\rho^2 \int_{\bar{t}}^{t+1} |\mathbf{v}(\zeta)|_{1,2}^2 d\zeta} \\ &\leq (\rho_1^2 + F) e^{\rho^2 \rho_1^2}, \quad \text{for all } t \geq t_1, \end{aligned}$$

which proves also property (b).  $\square$

**Remark 18** In the proof of the previous theorem the assumption of the boundedness of  $\Omega$  is crucial, in order to ensure the validity of Rellich's compactness theorem. However, a different approach, due to Rosa (see Theorem 3.2 in [73]) allows us to draw the same conclusion of Theorem 20 under the more general assumption that in  $\Omega$  the Poincaré inequality (18) holds. This happens whenever  $\Omega$  is contained in a strip of finite width (like in a flow in an infinite channel).

We shall now list some further properties of the global attractor  $\mathfrak{M}$ , for whose proofs we refer to the monographs [26, 59, 94]

**A. Smoothness** The restriction of the semigroup  $S_t$  to  $\mathfrak{M}$  can be extended to a group,  $\hat{S}_t$ , defined for all  $t \in (-\infty, \infty)$ . Therefore, the pair  $\{\mathfrak{M}, \hat{S}_t\}$  constitutes a *flow (dynamical system)*. This flow is as smooth as allowed by  $\mathbf{f}$  and  $\Omega$ . In particular, if  $\mathbf{f}$  and  $\Omega$  are of class  $C^\infty$ , then the solutions to (7)–(8)<sub>Hom</sub> belonging to  $\mathfrak{M}$  are of class  $C^\infty$  in space and time as well. Further significant regularity properties can be found in [23].

**B. Finite Dimensionality** Let  $X$  be a bounded set of a metric space and let  $N(X, \varepsilon)$  be the smallest number of balls of radius  $\varepsilon$  necessary to cover  $X$ . The non-negative (possibly infinite) number

$$d_f(X) = \limsup_{\varepsilon \rightarrow 0^+} \frac{\ln N(X, \varepsilon)}{\ln(1/\varepsilon)},$$

is called the *fractal dimension* of  $X$ . If  $X$  is closed with  $d_f(X) < \infty$ , then there exists a Lipschitz-continuous function,  $g : X \mapsto \mathbb{R}^m$ ,  $m > 2d_f(X)$ , possessing a Hölder-continuous inverse on  $g(X)$  (see Theorem 1.2 in [21]).

The fundamental result states that  $d_f(\mathfrak{M})$  is finite and that, moreover,

$$d_f(\mathfrak{M}) \leq c \|\mathbf{f}\|_2 / (\nu^2 C_P) := cG, \quad (112)$$

where  $c$  is a positive constant depending only on the “shape” of  $\Omega$ ; see [92]. The quantity  $G$  is non-dimensional (often called *Grashof number*). Consequently,  $\mathfrak{M}$  is (in particular) homeomorphic to a compact set of  $\mathbb{R}^m$ , with  $m = 2cG + 1$ . Notice that, in agreement with what conjectured by E. Hopf, (112) gives a rigorous estimate of how the dimension of  $\mathfrak{M}$  is expected to increase with the magnitude of the driving force. (Recall, however, that, as remarked previously, there are examples where  $d_f(\mathfrak{M}(G)) = 0$ , for all  $G > 0$ .)

**Open Question** Since  $\mathfrak{M}$  can be parametrized by a finite number of parameters, or, equivalently, it can be “smoothly” embedded in a finite-dimensional space, it is a natural question to ask whether or not one can construct a finite-dimensional dynamical system having a global attractor on which the dynamics is “equivalent” to the Navier–Stokes dynamics on  $\mathfrak{M}$ . This question, which is still *unresolved*, has led to the introduction of the idea of *inertial manifold* [27] and of the associated *approximate inertial manifold* [25], for whose definitions and detailed properties we refer to Chap. VIII in [94]; see, however, also [51].

**Further Questions Related to the Existence of the Global Attractor** In this section we shall address two further important aspects of the theory of attractors for the Navier–Stokes equations, namely, the three-dimensional case (in a bounded domain) and the case of a flow past an obstacle.

**A. Three-Dimensional Flow in a Bounded Domain** If we go through the first part of the proof of Theorem 20, we see that the assumption of planar flow has not been used. In fact, by the same token, we can still prove for three-dimensional flow the existence of an “absorbing set”, in the

sense that every solution departing from any bounded set of  $L^2_\sigma$ ,  $\mathcal{A}$ , will end up in the set  $\mathfrak{B}$  defined in (109), after a time  $t_0$ , dependent on  $\mathcal{A}$  and  $F$ . However, the difficulty in extending the results of Theorem 20 to three-dimensional flow, resides, fundamentally, in the lack of well-posedness of (7)–(8)<sub>Hom</sub> in the space  $L^2_\sigma(\Omega)$ ; see Subsect. “Uniqueness and continuous Dependence”. In order to overcome this situation, several strategies of attack have been proposed. One way is to make an *unproved* assumption on all possible solutions, which guarantees the existence of a semiflow on  $L^2_\sigma(\Omega)$ ; see [Chapter I] in [16]. In such a case, all the fundamental results proven for the 2D flow, continue to hold in 3D as well [16]. Another way is to weaken the definition of global attractor, by requiring the attractivity property in the weak topology of  $L^2$ ; see [Chapter III.3] in [26], and a third way is to generalize the definition of semiflow in such a way that the uniqueness property is no longer required; see [3,13,79].

**B. Flow Past an Obstacle** In this case, the relevant initial-boundary value to be investigated is (7)–(8) with  $\mathbf{f} \equiv \mathbf{v}_1 \equiv \mathbf{0}$ , endowed with the condition at infinity  $\lim_{|x| \rightarrow \infty} \mathbf{v}(x, t) = \mathbf{U}$ , where  $\mathbf{U} = U\mathbf{e}_1$  is a given, non-zero constant vector. For the reader’s convenience, we shall rewrite here this set of equations and put them in a suitable non-dimensional form:

$$\left. \begin{aligned} \frac{\partial \mathbf{v}}{\partial t} + \lambda \mathbf{v} \cdot \nabla \mathbf{v} &= \Delta \mathbf{v} - \nabla p \\ \operatorname{div} \mathbf{v} &= 0 \end{aligned} \right\} \quad \text{in } \Omega \times (0, \infty) \quad (113)$$

$$\mathbf{v}(x, t)|_{\partial\Omega} = \mathbf{0}, \quad \lim_{|x| \rightarrow \infty} \mathbf{v}(x, t) = \mathbf{e}_1, \quad t > 0;$$

$$\mathbf{v}(x, 0) = \mathbf{v}_0(x).$$

In (113),  $\lambda := |U|d/\nu$ , with  $d$  a length scale, is the appropriate Reynolds number which furnishes the magnitude of the driving mechanism.

As we observed in Remark 16, the least requirement on the spatial domain for the existence of a global attractor in a two-dimensional flow, is that there holds Poincaré’s inequality (18). Since this inequality is no longer valid in an exterior domain, the problem of the existence of an attractor for a flow past an obstacle is, basically, unresolved. Actually, the situation is even more complicated than what we just described. In fact, from Theorem 9 and from the considerations developed after it, we know that there is  $\lambda_c > 0$  such that if  $\lambda < \lambda_c$ , the corresponding boundary-value problem has one and only one solution,  $(\bar{\mathbf{v}}, \bar{p})$ , in a suitable function class. Now, it is not known if this solution is attracting. More precisely, *in the two-dimensional flow, it is not known if, for sufficiently small  $\lambda$ , solutions to*

(113), *defined in an appropriate class, tend, as  $t \rightarrow \infty$ , to the only corresponding steady solution.*

In the three-dimensional case, the situation is slightly better, but still, the question of the existence of an attractor is completely open. We would like to go into more detail about this point. We begin to observe that there is  $\lambda_0 > 0$  such that if  $\lambda < \lambda_0$ , for any given  $\mathbf{v}_0$  with  $(\mathbf{v}_0 - \mathbf{e}_1) \in L^3_\sigma(\Omega)$ , problem (113) has one and only one (smooth) solution that tends to the (uniquely determined) corresponding steady-state solution,  $(\bar{\mathbf{v}}, \bar{p})$ . In particular

$$\lim_{t \rightarrow \infty} \|\mathbf{v}(t) - \bar{\mathbf{v}}\|_3 = 0; \quad (114)$$

see [41]. The fundamental question that stays open is then that of investigating the behavior of solutions to (113) for large  $t$ , when  $\lambda > \lambda_0$ . As a matter of fact, *it is not known whether there exists a norm with respect to which solutions to (113), in a suitable class, remain bounded uniformly in time, for all  $\lambda > 0$ .* In this respect, it is readily seen that, unlike the bounded domain situation, solutions to (113), in general, can not be bounded in  $L^2(\Omega)$ , uniformly in time, *even when  $\lambda < \lambda_0$ .* This means that the kinetic energy associated to the motion described by (113) has to grow unbounded for large times. To see this, assume  $\lambda < \lambda_0$  and that there exists  $K > 0$ , independent of  $t$ , such that

$$\|\mathbf{v}(t) - \mathbf{e}_1\|_2 \leq K, \quad (115)$$

where  $\mathbf{v}$  is a solution to (113). Then, we can find an unbounded sequence,  $\{t_m\}$ , and an element  $\bar{\mathbf{w}} \in L^2_\sigma(\Omega)$  (possibly depending on the sequence) such that

$$\lim_{m \rightarrow \infty} (\mathbf{v}(t_m) - \mathbf{e}_1, \boldsymbol{\varphi}) = (\bar{\mathbf{w}}, \boldsymbol{\varphi}), \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega). \quad (116)$$

By (114) and (116) we thus must have  $\bar{\mathbf{w}} = \bar{\mathbf{v}} - \mathbf{e}_1$ , which in turn implies  $(\bar{\mathbf{v}} - \mathbf{e}_1) \in L^2_\sigma(\Omega)$ , which, from Theorem 6, we know to be impossible. Consequently, (115) can not be true. Thus, the basic open question is whether or not there exists a function space,  $Y$ , where the solution  $\mathbf{v}(t) - \mathbf{e}_1$  remains *uniformly* bounded in  $t \in (0, \infty)$ , for *all*  $\lambda > 0$ . (The bound, of course, may depend on  $\lambda$ .) The above considerations along with Theorem 6 suggest then that a plausible candidate for  $Y$  is  $L^q_\sigma(\Omega)$ , for *some*  $q > 2$ . However, the proof of this property for  $q \geq 3$  appears to be overwhelmingly challenging because, in view of Theorem 18 and Remark 8, it would be closely related to the existence of a global, regular solution. Nevertheless, one could investigate the validity of the following weaker property

$$\|\mathbf{v}(t) - \mathbf{e}_1\|_q \leq K_1, \quad \text{for some } q \in (2, 3), \quad (117)$$

where  $K_1$  is independent of  $t \in (0, \infty)$ . Of course, the requirement is that (117) holds for all  $\lambda > 0$  and for all corresponding solutions. It is worth emphasizing that the proof of (117) would be of “no harm” to the outstanding global regularity Problem 2, since, according to the available regularity criteria for weak solutions that we discussed in Sect. “Less Regular Solutions and Partial Regularity Results in 3D”, the corresponding solutions, while global in time, will still be weak, even though more regular than those described in Theorem 16. However, notwithstanding its plausibility and “harmlessness”, the property (117) appears to be very difficult to establish.

### Future Directions

The fundamental open questions that we have pointed out throughout this work constitute as many topics for future investigation. Actually, it is commonly believed that the answer to most of these questions (in the affirmative or in the negative) will probably shed an entirely new light not only on the mathematical theory of the Navier–Stokes equations but also on other disciplines of applied mathematics.

### Acknowledgment

This work was partially supported by the National Science Foundation, Grants DMS-0404834 and DMS-0707281.

### Bibliography

- Amick CJ (1984) Existence of Solutions to the Nonhomogeneous Steady Navier–Stokes Equations. *Indiana Univ Math J* 33:817–830
- Amick CJ (1988) On Leray’s Problem of Steady Navier–Stokes Flow Past a Body in the Plane. *Acta Math* 161:71–130
- Ball JM (1997) Continuity Properties and Global Attractors of Generalized Semiflows and the Navier–Stokes Equations. *J Nonlinear Sci* 7:475–502
- Batchelor GK (1981) *An Introduction to Fluid Mechanics*. Cambridge University Press, Cambridge
- Beirão da Veiga H (1985) On the Construction of Suitable Weak Solutions to the Navier–Stokes Equations via a General Approximation Theorem. *J Math Pures Appl* 64:321–334
- Beirão da Veiga H (1995) A New Regularity Class for the Navier–Stokes Equations in  $R^n$ . *Chinese Ann Math Ser B* 16:407–412
- Berger MS (1977) *Nonlinearity and Functional Analysis. Lectures on Nonlinear Problems in Mathematical Analysis*. Academic Press, New York
- Berselli LC, Galdi GP (2002) Regularity Criteria Involving the Pressure for the Weak Solutions to the Navier–Stokes Equations. *Proc Amer Math Soc* 130:3585–3595
- Brown RM, Shen Z (1995) Estimates for the Stokes Operator in Lipschitz Domains. *Indiana Univ Math J* 44:1183–1206
- Caccioppoli R (1936) *Sulle Corrispondenze Funzionali Inverse Diramate: Teoria Generale ed Applicazioni al Problema di Plateau*. *Rend Accad Lincei* 24:258–263; 416–421
- Caffarelli L, Kohn R, Nirenberg L (1982) Partial Regularity of Suitable Weak Solutions of the Navier–Stokes Equations. *Comm Pure Appl Math* 35:771–831
- Cattabriga L (1961) Su un Problema al Contorno Relativo al Sistema di Equazioni di Stokes. *Rend Sem Mat Padova* 31:308–340
- Cheskidov A, Foias C (2006) On Global Attractors of the 3D Navier–Stokes Equations. *J Diff Eq* 231:714–754
- Constantin P (1990) Navier–Stokes Equations and Area of Interfaces. *Comm Math Phys* 129:241–266
- Constantin P, Fefferman C (1993) Direction of Vorticity and the Problem of Global Regularity for the Navier–Stokes Equations. *Indiana Univ Math J* 42:775–789
- Constantin P, Foias C, Temam R (1985) Attractors Representing Turbulent Flows. *Mem Amer Math Soc* 53 no. 314 vii, pp 67
- Darrigol O (2002) Between Hydrodynamics and Elasticity Theory: The First Five Births of the Navier–Stokes Equation. *Arch Hist Exact Sci* 56:95–150
- Escuriaza L, Seregin G, Sverák V (2003) Backward Uniqueness for Parabolic Equations. *Arch Ration Mech Anal* 169:147–157
- Federer H (1969) *Geometric Measure Theory. Die Grundlehren der mathematischen Wissenschaften*, vol 153. Springer, New York
- Finn R, Smith DR (1967) On the stationary solution of the Navier–Stokes equations in two dimensions. *Arch Rational Mech Anal* 25:26–39
- Foias C, Olson EJ (1996) Finite Fractal Dimension and Hölder–Lipschitz Parametrization. *Indiana Univ Math J* 45:603–616
- Foias C, Temam R (1977) Structure of the Set of Stationary Solutions of the Navier–Stokes Equations. *Comm Pure Appl Math* 30:149–164
- Foias C, Temam R (1989) Gevrey Class Regularity for the Solutions of the Navier–Stokes Equations. *J Func Anal* 87:359–69
- Foias C, Guillopé C, Temam R (1981) New A Priori Estimates for Navier–Stokes Equations in Dimension 3. *Comm Partial Diff Equ* 6:329–359
- Foias C, Manley OP, Temam R (1988) Modelling of the Interaction of Small and Large Eddies in Two-Dimensional Turbulent Flows. *RAIRO Modél Math Anal Numér* 22:93–118
- Foias C, Manley OP, Rosa R, Temam R (2001) *Navier–Stokes Equations and Turbulence. Encyclopedia of Mathematics and its Applications*, 83. Cambridge University Press, Cambridge
- Foias C, Sell GR, Temam R (1988) Inertial Manifolds for Nonlinear Evolutionary Equations. *J Diff Eq* 73:309–353
- Fujita H (1961) On the Existence and Regularity of the Steady-State Solutions of the Navier–Stokes Equation. *J Fac Sci Univ Tokyo* 9:59–102
- Fujita H (1998) On stationary solutions to Navier–Stokes equation in symmetric plane domains under general outflow condition. In Salvi R (ed) *Navier–Stokes Equations: Theory and Numerical Methods*. Pitman Res Notes Math Ser 388:16–30
- Galdi GP (1993) Existence and Uniqueness at Low Reynolds Number of Stationary Plane Flow of a Viscous Fluid in Exterior Domains. In: Galdi GP, Necas J (eds) *Recent Developments in Theoretical Fluid Mechanics*. Pitman Res Notes Math Ser Longman Sci Tech, Harlow 291:1–33



31. Galdi GP (1998) An Introduction to the Mathematical Theory of the Navier–Stokes Equations. vol I. Linearized Steady Problems. Springer Tracts in Natural Philosophy, 38. Springer, New York (revised Edition)
32. Galdi GP (1998) An Introduction to the Mathematical Theory of the Navier–Stokes Equations. vol II. Nonlinear Steady Problems. Springer Tracts in Natural Philosophy, 39. Springer, New York (revised Edition)
33. Galdi GP (1999) On the Existence of Symmetric Steady-State Solutions to the Plane Exterior Navier–Stokes Problem for Arbitrary Large Reynolds Number. In: Maremonti P (ed) Advances in fluid dynamics. Quad Mat Aracne Rome 4:1–25
34. Galdi GP (2000) An Introduction to the Navier–Stokes Initial-Boundary Value Problem. In: Galdi GP, Heywood JG, Rannacher R (eds) Fundamental Directions in Mathematical Fluid Mechanics. Adv Math Fluid Mech Birkhäuser, Basel 1:1–70
35. Galdi GP (2004) Stationary Navier–Stokes Problem in a Two-Dimensional Exterior Domain. In: Chipot M, Quittner P (eds) Stationary partial differential equations. Handb Diff Equ, North-Holland, Amsterdam 1:71–155
36. Galdi GP (2007) Further Properties of Steady-State Solutions to the Navier–Stokes Problem Past a Threedimensional Obstacle. J Math Phys 48:1–43
37. Galdi GP (2007) Some Mathematical Properties of the Steady-State Navier–Stokes Problem Past a Three-Dimensional Obstacle. RWTH Aachen Institut für Mathematik, Report no. 17
38. Galdi GP, Maremonti P (1988) Regularity of Weak Solutions of the Navier–Stokes System in Arbitrary Domains. Ann Univ Ferrara Sez VII 34:59–73
39. Galdi GP, Padula M (1990) A new approach to energy theory in the stability of fluid motion. Arch Rational Mech Anal 110:187–286
40. Galdi GP, Rabier PJ (1999) Functional Properties of the Navier–Stokes Operator and Bifurcation of Stationary Solutions: Planar Exterior Domains. In: Escher J, Simonett G (eds) Topics in Nonlinear Analysis. Progr Nonlinear Diff Equ Appl. Birkhäuser, Basel 35:273–303
41. Galdi GP, Heywood JG, Shibata Y (1997) On the Global Existence and Convergence to Steady State of Navier–Stokes Flow Past an Obstacle that is Started from Rest. Arch Rational Mech Anal 138:307–318
42. Galdi GP, Robertson AM, Rannacher R, Turek S (2007) Hemodynamical Flows: Modeling, Analysis and Simulation. Oberwolfach Seminar Series vol 35. Birkhäuser
43. Giga Y, Sohr H (1991) Abstract  $L^p$  Estimates for the Cauchy Problem with Applications to the Navier–Stokes Equations in Exterior Domains. J Funct Anal 102:72–94
44. Gilbarg D, Weinberger HF (1974) Asymptotic Properties of Leray's Solution of the Stationary Two-Dimensional Navier–Stokes Equations. Russian Math Surveys 29:109–123
45. Gilbarg D, Weinberger HF (1978) Asymptotic Properties of Steady Plane Solutions of the Navier–Stokes Equations with Bounded Dirichlet. Integral Ann Scuola Norm Sup Pisa 5:381–404
46. Gohberg I, Goldberg S, Kaashoek MA (1990) Classes of Linear Operators: I. Operator Theory. Advances and Applications. vol 49. Birkhäuser, Basel
47. Guermond JL (2007) Faedo–Galerkin Weak Solutions of the Navier–Stokes Equations with Dirichlet Boundary Conditions are Suitable. J Math Pures Appl 88:87–106
48. Gustafson S, Kang K, Tsai TP (2006) Regularity Criteria for Suitable Weak Solutions of the Navier–Stokes Equations Near the Boundary. J Diff Equ 226:594–618
49. Gustafson S, Kang K, Tsai TP (2007) Interior Regularity Criteria for Suitable Weak Solutions of the Navier–Stokes Equations. Comm Math Phys 273:161–176
50. Heywood JG (1980) The Navier–Stokes Equations: on the Existence, Regularity and Decay of Solutions. Indiana Univ Math J 29:639–681
51. Heywood JG, Rannacher R (1993) On the Question of Turbulence Modeling by Approximate Inertial Manifolds and the Nonlinear Galerkin Method. SIAM J Numer Anal 30:1603–1621
52. Hopf E (1941) Ein Allgemeiner Endlichkeitsatz der Hydrodynamik. Math Annalen 117:764–775
53. Hopf E (1948) A Mathematical Example Displaying Features of Turbulence. Comm Pure Appl Math 1:303–322
54. Hopf E (1951) Über die Anfangswertaufgabe für die hydrodynamischen Grundgleichungen. Math Nachr 4:213–231
55. Kozono H, Sohr H (2000) Remark on Uniqueness of Weak Solutions to the Navier–Stokes Equations. Analysis 16:255–271
56. Kozono H, Taniuchi Y (2000) Bilinear Estimates in BMO and the Navier–Stokes Equations. Math Z 235:173–94
57. Kozono H, Yatsu N (2004) Extension Criterion via Two-Components of Vorticity on Strong Solutions to the 3D Navier–Stokes Equations. Math Z 246:55–68
58. Ladyzhenskaya OA (1963) The mathematical theory of viscous incompressible flow. Revised English edition. Gordon and Breach Science, New York-London
59. Ladyzhenskaya OA (1991) Attractors for Semigroups and Evolution Equations. Lezioni Lincee. Cambridge University Press, Cambridge
60. Ladyzhenskaya OA, Seregin GA (1999) On Partial Regularity of Suitable Weak Solutions to the Three-Dimensional Navier–Stokes Equations. J Math Fluid Mech 1:356–387
61. Leray J (1933) Etude de Diverses Équations Intégrales non Linéaires et de Quelques Problèmes que Pose l'Hydrodynamique. J Math Pures Appl 12:1–82
62. Leray J (1934) Sur le Mouvement d'un Liquide Visqueux Emplissant l'Espace. Acta Math 63:193–248
63. Leray J (1936) Les Problèmes non Linéaires. Enseignement Math 35:139–151
64. Lin F (1998) A New Proof of the Caffarelli–Kohn–Nirenberg Theorem. Comm. Pure Appl Math 51:241–257
65. Marchioro C (1986) An Example of Absence of Turbulence for any Reynolds Number. Comm Math Phys 105:99–106
66. Maremonti P (1998) Some Interpolation Inequalities Involving Stokes Operator and First Order Derivatives. Ann Mat Pura Appl 175:59–91
67. Málek J, Padula M, Ruzicka M (1995) A Note on Derivative Estimates for a Hopf Solution to the Navier–Stokes System in a Three-Dimensional Cube. In: Sequeira A (ed) Navier–Stokes Equations and Related Nonlinear Problems. Plenum, New York 141–146
68. Natarajan R, Acrivos A (1993) The Instability of the Steady Flow Past Spheres and Disks. J Fluid Mech 254:323–342
69. Neustupa J, Novotný A, Penel P (2002) An Interior Regularity of a Weak Solution to the Navier–Stokes Equations in Dependence on one Component of Velocity. In: Galdi GP, Rannacher R (eds) Topics in mathematical fluid mechanics. Quad Mat Aracne, Rome 10:163–183

70. Nirenberg L (1959) On Elliptic Partial Differential Equations. *Ann Scuola Norm Sup Pisa* 13:115–162
71. Prodi G (1959) Un Teorema di Unicità per le Equazioni di Navier–Stokes. *Ann Mat Pura Appl* 48:173–182
72. Prodi G (1962) Teoremi di Tipo Locale per il Sistema di Navier–Stokes e Stabilità delle Soluzioni Stazionarie. *Rend Sem Mat Univ Padova* 32:374–397
73. Rosa R (1998) The Global Attractor for the 2D Navier–Stokes Flow on Some Unbounded Domains. *Nonlinear Anal* 32:71–85
74. Ruelle D, Takens F (1971) On the Nature of Turbulence. *Comm Math Phys* 20:167–192
75. Sather J (1963) The Initial Boundary Value Problem for the Navier–Stokes Equations in Regions with Moving Boundaries. Ph.D. thesis, University of Minnesota
76. Scheffer V (1976) Partial Regularity of Solutions to the Navier–Stokes Equations. *Pacific J Math* 66:535–552
77. Scheffer V (1977) Hausdorff Measure and the Navier–Stokes Equations. *Comm Math Phys* 55:97–112
78. Scheffer V (1980) The Navier–Stokes Equations on a Bounded Domain. *Comm Math Phys* 73:1–42
79. Sell GR (1996) Global Attractors for the Three-Dimensional Navier–Stokes Equations. *J Dynam Diff Eq* 8:1–33
80. Seregin G, Sverák V (2002) The Navier–Stokes Equations and Backward Uniqueness. In: Birman MS, Hildebrandt S, Solonnikov VA, Uraltseva NN (eds) *Nonlinear problems in mathematical physics and related topics, II*. *Int Math Ser (N. Y.)* Kluwer/Plenum, New York 2:353–366
81. Seregin GA, Shilkin TN, Solonnikov VA (2006) Boundary Partial Regularity for the Navier–Stokes Equations. *J Math Sci* 132:339–358
82. Serrin JB (1962) On the Interior Regularity of Weak Solutions of the Navier–Stokes Equations. *Arch Rational Mech Anal* 9:187–195
83. Serrin JB (1963) The Initial Value Problem for the Navier–Stokes Equations. In: Langer RE (ed) *Nonlinear Problems*. University of Wisconsin Press, pp. 69–98
84. Smale S (1965) An Infinite Dimensional Version of Sard’s Theorem. *Amer J Math* 87:861–866
85. Smale S (1967) Differentiable Dynamical Systems. *Bull Amer Math Soc* 73:747–817
86. Sohr H (2001) *The Navier–Stokes Equations. An Elementary Functional Analytic Approach*. Birkhäuser, Basel
87. Sohr H, von Wahl W (1984) On the Singular Set and the Uniqueness of Weak Solutions of the Navier–Stokes Equations. *Manuscripta Math* 49:27–59
88. Stein EM (1970) *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, Princeton
89. Stokes GG (1851) On the Effect of the Internal Friction of Fluids on the Motion of Pendulums. *Trans Cambridge Phil Soc* 9:8–106
90. Takeshita A (1993) A Remark on Leray’s Inequality. *Pacific J Math* 157:151–158
91. Taneda S (1956) Experimental Investigation of the Wake Behind a Sphere at Low Reynolds Numbers. *J Phys Soc Japan* 11:1104–1111
92. Temam R (1986) Infinite-Dimensional Dynamical Systems in Fluid Mechanics. In: Browder F (ed) *Nonlinear Functional Analysis and its Applications, Part 2*. *Proc. Sympos. Pure Math Amer. Math Soc.* Providence, RI 45:431–445
93. Temam R (1995) *Navier–Stokes Equations and Nonlinear Functional Analysis*. CBMS-NSF Regional Conference Series in Applied Mathematics, 66
94. Temam R (1997) *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*. Second edition. *Applied Mathematical Sciences*, 68. Springer, New York
95. Tomboulides AG, Orszag SA (2000) Numerical Investigation of Transitional and Weak Turbulent Flow Past a Sphere. *J Fluid Mech* 416:45–73
96. Vorovich II, Youdovic VI (1961) Stationary Flow of a Viscous Incompressible Fluid. *Mat Sb* 53:393–428 (in Russian)
97. Wu JS, Faeth GM (1993) Sphere Wakes in Still Surroundings at Intermediate Reynolds Numbers. *AIAA J* 31:1448–1460
98. Xie W (1997) Sharp Sobolev Interpolation Inequalities for the Stokes Operator. *Diff Inte Equ* 10:393–399
99. Zeidler E (1988) *Nonlinear Functional Analysis and Applications: Application to Mathematical Physics*. Springer, New York
100. Zeidler E (1995) *Applied Functional Analysis: Main Principles and their Applications*. *Applied Math Sci.*, Springer, vol 109

## n-Body Problem and Choreographies

SUSANNA TERRACINI

Dipartimento di Matematica e Applicazioni,  
Università di Milano Bicocca, Milano, Italia

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Simple Choreographies and Relative Equilibria](#)

[Symmetry Groups and Equivariant Orbits](#)

[The 3-Body Problem](#)

[Minimizing Properties of Simple Choreographies](#)

[Generalized Orbits and Singularities](#)

[Asymptotic Estimates at Collisions](#)

[Absence of Collision for Locally Minimal Paths](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Central configurations** Are the critical points of the potential constrained on the unitary moment of inertia ellipsoid. Central configurations are associated to particular solutions to the *n*-body problem: the *relative equilibrium* and the *homographic* motions defined in Definition 1.

**Choreographical solution** A choreographical solution of the *n*-body problem is a solution such that the particles move on the same curve, exchanging their positions after a fixed time. This property can be re-

garded as a symmetry of the trajectory. This notion finds a natural generalization in that of *G-equivariant trajectory* defined in Definition 2 for a given group of symmetries- $G$ . The *G-equivariant minimization technique* consists in seeking action minimizing trajectories among all  $G$ -equivariant paths.

**Collision and singularities** When a trajectory can not be extended beyond a certain time  $b$  we say that a *singularity* occurs. Singularities can be *collisions* if the solution admits a limit configuration as  $t \rightarrow b$ . In such a case we term  $b$  a *collision instant*.

**n-Body problem** The  $n$ -body problem is the system of differential equations (1) associated with suitable initial or boundary value data. A *solution* or *trajectory* is a doubly differentiable path  $q(t) = (q_1(t), \dots, q_n(t))$  satisfying (1) for all  $t$ . The weaker notion of *generalized solution* is defined in Definition 5 applies to trajectories found by variational methods.

**Variational approach** The variational approach to the  $n$ -body problem consists in looking at trajectories as critical points of the *action functional* defined in (4). Such critical points can be (local) *minimizers*, or *constrained minimizers* or *mountain pass*, or other type.

### Definition of the Subject

The motion of  $n$ -point particles of positions  $x_i(t) \in \mathbb{R}^3$  and masses  $m_i > 0$ , interacting in accordance with Newton's law of gravitation, satisfies the system of *differential equations*:

$$-m_i \ddot{x}_i(t) = G \sum_{j \neq i, j=1}^n m_i m_j \frac{x_i - x_j}{|x_i - x_j|^3}, \quad i = 1, \dots, n, \quad t \in \mathbb{R}. \quad (1)$$

The *n-body problem* consists in solving Eq. (1) associated with initial or boundary conditions.

A *simple choreography* is a periodic solution to the  $n$ -body problem Eq. (1) where the bodies lie on the same curve and exchange their mutual positions after a fixed time, namely, there exists a function  $x: \mathbb{R} \rightarrow \mathbb{R}$  such that

$$x_i(t) = x(t + (i-1)\tau), \quad i = 1, \dots, n, \quad t \in \mathbb{R}, \quad (2)$$

where  $\tau = 2\pi/n$ .

### Introduction

The *two-body problem* can be reduced, by the conservation of the linear momentum, to the *one center Kepler problem*

and can be completely solved either by exploiting the conservation laws (angular momentum, energy and the Lenz vector), or by performing the Levi-Civita change of coordinates reducing the problem to that of an harmonic oscillator [58].

The three-body problem is much more complicated than the two-body and can not be solved in a simple way. A major study of the Earth-Moon-Sun system was undertaken by Delaunay in his *La Théorie du mouvement de la lune*. In the *restricted three-body problem*, the mass of one of the bodies is negligible; the circular restricted three-body problem is the special case in which two of the bodies are in circular orbits and was worked on extensively by many famous mathematicians and physicists, notably Lagrange in the 18th century, Poincaré at the end of the 19th century and Moser in the 20th century. Poincaré's work on the restricted three-body problem was the foundation of deterministic chaos theory.

A very basic – still fundamental – question concerns the real number of degrees of freedom of the  $n$ -body problem. As the motion of each point particle is represented by 3-dimensional vectors, the  $n$ -body problem has  $3n$  degrees of freedom and hence it is  $6n$ -dimensional. First integrals are: the center of mass, the linear momentum, the angular momentum, and the energy. Hence there are 10 independent algebraic integrals. This allows the reduction of variables to  $6n - 10$ . It was proved in 1887 by Bruns that these are the only linearly independent integrals of the  $n$ -body problem, which are algebraic with respect to phase and time variables. This theorem was later generalised by Poincaré.

A second very natural problem is whether there exists a power series expressing – if not every – at least a large and relevant class of trajectories for the  $n$ -body problem. In 1912, after pioneering works of Mittag-Leffler and Levi-Civita, Sundman proved the existence of a series solution in powers of  $t^{1/3}$  for the 3-body problem. This series is convergent for all real  $t$ , except for those initial data which correspond to *vanishing angular momentum*. These initial data have Lebesgue measure zero and therefore are not generic. An important issue in proving this result is the fact that the radius of convergence for this series is determined by the distance to the nearest singularity. Since then, the study of singularities became the main point of interest in the study of the  $n$ -body problem. Sundman's result was later generalised to the case of  $n > 3$  bodies by Q. Wang in [88]. However, the rate and domain of convergence of this series are so limited to make it hardly applicable to practical and theoretical purposes.

Finally, the  $n$ -body problem can be faced from the point of view of the theory of perturbations and represents

both its starting point and its most relevant application. Delaunay and Poincaré described the spatial three-body problem as a four-dimensional Hamiltonian system and already encountered some trajectories featuring a chaotic behavior. When one of the bodies is much heavier than the other two (a system of one “star” and two “planets”), one can neglect the interaction between the small planets; hence the system can be seen as a perturbation of two decoupled two-body problems whose motions are known to be all periodic from Kepler’s laws. In the planetary three-body problem, hence, two harmonic oscillators interact nonlinearly through the perturbation. Resonances between the two oscillators can be held responsible of high sensitivity with respect to initial data and other chaotic features. A natural question regards the coexistence of the irregular trajectories with regular (periodic or quasi-periodic) ones. A modern approach to the problem of stability of solutions to nearly integrable systems goes through the application of Kolmogorov–Arnold–Moser (KAM) Theorem [8,56], whose main object indeed is indeed the persistence, under perturbations, of invariant tori.

The  $n$ -body problem is paradigmatic of any complex system of many interacting objects, and it can neither be solved nor it can be simplified in an efficient way. A possible starting point for its analysis is to seek selected trajectories whose motion is particularly simple, in the sense that it repeats after a fixed period: the periodic solutions. Following Poincaré in his *Méthodes Nouvelles de la Mécanique Céleste*, tome I, 1892,

“... D’ailleurs, ce qui nous rends ces solutions périodiques si précieuses, c’est qu’elles sont, pour ainsi dire, la seule brèche par où nous puissions essayer de pénétrer dans une place jusqu’ici réputée inabordable.”

Indeed, just before, Poincaré conjectured that periodic trajectories are dense in the phase space:

“... Voici un fait que je n’ai pas pu démontrer rigoureusement, mais qui me paraît pourtant très vraisemblable. Étant données des équations de la forme définie dans le n.13<sup>1</sup> et une solution quelconque de ces équations, on peut toujours trouver une solution périodique (dont la période peut, il est vrai, être très longue), telle que la différence entre les deux solutions soit aussi petite qu’on le veut, pendant un temps aussi long qu’on le veut.”

<sup>1</sup>Formula N. 13 quoted by Poincaré is Hamilton equation and covers our class of Dynamical Systems Eq. (3)

## Singular Hamiltonian Systems

From an abstract point of view, the  $n$ -body problem is a Hamiltonian System of the form

$$m_i \ddot{x}_i = \frac{\partial U}{\partial x_i}(t, x), \quad i = 1, \dots, n, \quad (3)$$

where the forces  $\frac{\partial U}{\partial x_i}$  are undefined on a singular set  $\Delta$ , the set of collisions between two or more particles in the  $n$ -body problem. Such singularities play a fundamental role in the phase portrait (see, e. g. [43]) and strongly influence the global orbit structure, as they can be held responsible, among others, of the presence of *chaotic motions* (see, e. g. [39]) and of *motions becoming unbounded in a finite time* [64,90].

Two are the major steps in the analysis of the impact of the singularities in the  $n$ -body problem: the first consists in performing the asymptotic analysis along a single collision (total or partial) trajectory and goes back, in the classical case, to the works by Sundman ([84]), Wintner ([89]) and, in more recent years by Sperling, Pollard, Saari and other authors (see for instance [40,47,75,76,79,83]). The second step consists in blowing-up the singularity by a suitable change of coordinates introduced by McGehee in [65] and replacing it by an invariant boundary – the collision manifold – where the flow can be extended in a smooth manner. It turns out that, in many interesting applications, the flow on the collision manifold has a simple structure: it is a gradient-like, Morse–Smale flow featuring a few stationary points and heteroclinic connections (see, for instance, the surveys [39,68]). The analysis of the extended flow allows us to obtain a full picture of the behavior of solutions near the singularity, despite the flow fails to be fully regularizable (except for binary collisions).

## Simple Choreographies and Relative Equilibria

A possible starting point for the study of the  $n$ -body problem is to find selected trajectories which are particularly simple, when regarded from some point of view. Examples of such particular solutions are the collinear periodic orbits (found in 1767 by Euler), in which three bodies of any masses move such that they oscillate along a rotation line, and the Lagrange triangular solutions, where the bodies lie at the vertices of a rotating equilateral triangle that shrinks and expands periodically, discovered in 1772. Both these trajectories are stationary in a rotating frame. A second remarkable class of trajectories – the choreographies – can be found when the masses are all equal, by exploiting the fact that particles can be interchanged without changing the structure of the system.



Among all periodic solutions of the planar 3-body problem, the relative equilibrium motions – the equilateral Lagrange and the collinear Euler–Moulton solutions – are definitely the simplest and most known. In general such simple periodic motions exist for any number of bodies.

**Definition 1** A *relative equilibrium* trajectory is a solution of (1) whose configuration remains constant in a rotating frame. A *homographic* trajectory is a solution of (1) whose configuration remains constant up to homotheties.

The normalized configurations of such trajectories are named *central* and are the critical points of the potential

$$U(x) = \sum_{i < j} \frac{m_i m_j}{|x_i - x_j|},$$

constrained to the ellipsoid of unitary momentum of inertia:

$$I(x) = \sum_i m_i |x_i|^2 = 1.$$

Relative equilibria feature an evident symmetry ( $SO(2)$  and  $O(2)$  respectively), that is, they are equivariant with respect to the symmetry group of dimension 1 acting as  $SO(2)$  (resp.  $O(2)$ ) on the time circle and on the plane, and trivially on the set of indexes  $\{1, 2, 3\}$ . In fact, they are minimizers of the Lagrangian action functional in the space of all loops having their same symmetry group. Therefore, generally speaking, *G-equivariant minimizers for the action functional (given a symmetry group G) can be thought as the natural generalization of relative equilibrium motions.*

This perspective has known a wide popularity in the recent literature and has produced a new boost in the study of periodic trajectories in the  $n$ -body problem; the recent proof of the existence of the Chenciner–Montgomery eight-shaped orbit is emblematic of this renewed interest (see [7,20,22,25,29,52] and the major part of our bibliographical references). In all these papers, periodic and quasi-periodic solutions of the  $n$ -body problem can be found as critical points of the Lagrangian action functional restricted to suitable spaces of symmetric paths.

### Basic Definitions and Notations

Let us consider  $n$  point particles with masses  $m_1, m_2, \dots, m_n$  and positions  $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ , with  $d \geq 2$ . We denote by  $\mathcal{X}$  the space of configurations with center of mass in 0, and by  $\hat{\mathcal{X}} = \mathcal{X} \setminus \Delta$  the set of collision-free configurations (collision means  $x_i = x_j$  for some  $i \neq j$ ). On the configuration space we define the homogeneous (Newton)

potential of degree  $-\alpha < 0$ :

$$U(x) = \sum_{i < j} U_{i,j}(|x_i - x_j|), \quad U_{i,j}(|x_i - x_j|) \simeq \frac{m_i m_j}{|x_i - x_j|^\alpha}.$$

In many cases one can simply require the  $U_{i,j}$  to be asymptotically homogeneous at the singularity. Furthermore, a major part of our analysis can be extended to logarithmic potentials. Relative equilibria correspond to those configurations (termed central) which are critical for the restriction of the potential  $U$  to the ellipsoid  $I = \lambda$  where  $I$  denotes the momentum of inertia:

$$I(x) = \sum_i m_i |x_i|^2.$$

On collisions the potential  $U = +\infty$ . We are interested in (relative) periodic (such that  $\forall t: x(t+T) = x(t)$ ) solutions to the system of differential equations:

$$m_i \ddot{x}_i = \frac{\partial U}{\partial x_i}.$$

We associate with the equation the Lagrangian integrand

$$L(x, \dot{x}) = L = K + U = \sum_i \frac{1}{2} m_i |\dot{x}_i|^2 + \sum_{i < j} U_{i,j}(|x_i - x_j|)$$

and the action functional:

$$\mathcal{A}(x) = \int_0^T L(x(t), \dot{x}(t)) dt. \quad (4)$$

Sometimes it will be preferable to consider the problem in a frame rotating uniformly about the vertical axis, with an angular speed  $\omega$ ; the corresponding action  $\mathcal{A}^\omega$  then contains a gyroscopic term, associated with Coriolis force.

We shall seek periodic solutions as critical points of the action functional on the Sobolev space of  $T$ -periodic trajectories:  $\Lambda = H^1(\mathbb{T}, \mathcal{X})$  or, to be more precise, of the action constrained on suitable linear subspaces  $\Lambda_0 \subset \Lambda$ .

Two are the major difficulties to be faced in following the variational approach; the first is due to the *lack of coercivity* (or of Palais–Smale) due to the vanishing at infinity of the force fields: indeed sequences of almost-critical points (such as minimizing sequences) may very well diverge. Furthermore, as the potential  $U$  is singular on collisions, minimizers or other critical points can a priori be *collision trajectories*. Compactness can be successfully recovered by the symmetry of the problem, as we are going to explain here below. We shall expose in the last part of this article some of the strategies which has been developed in order to overcome the problem of collisions.



### Symmetry Groups and Equivariant Orbits

We can generalize the concept of relative equilibria as follows: we impose the permutation of the positions of the particles, after a given time, up to isometries of the space. This gives rise to a class of symmetric trajectories (the generalized choreographies). It is worthwhile noticing that these trajectories arise as a common feature of any system of interacting objects, regardless whether they come from models in Celestial mechanics or not. It applies to atoms, or molecules, galaxies or whatever system, provided the objects can be freely interchanged. Let us start by introducing some basics concepts and definitions from [52]. Let  $G$  be a finite group endowed with:

- an orthogonal representation of dimension 2  $\tau: G \rightarrow O(2)$  (on cyclic time  $\mathbb{T} \cong S^1$ ),
- an orthogonal representation (on the euclidean space  $\mathbb{R}^d$ )  $\rho: G \rightarrow O(d)$ ,
- and an homomorphism on the symmetric group on  $n$  elements ( $\mathbf{n} = \{1, 2, \dots, n\}$ )  $\sigma: G \rightarrow \Sigma_n$ .

Then  $G$  acts on time (translation and reversal)  $\mathbb{T}$  via  $\tau$  and it acts on the configuration space  $X$  via  $\rho$  and  $\sigma$  in the following way

$$\forall i = 1 \dots n: (gx)_i = \rho(g)x_{\sigma(g)^{-1}(i)}.$$

As a consequence we have an action on the space of trajectories:

**Definition 2** A continuous function  $x(t) = (x_1(t), \dots, x_n(t))$  is  $G$ -equivariant if

$$\forall g \in G: x(gt) = (gx)(t).$$

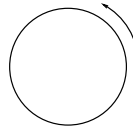
The linear subspace  $\Lambda_0 = \Lambda^G \subset \Lambda$  denotes the set of periodic trajectories in  $\Lambda$  which are equivariant with respect to the  $G$ -action:

The Palais Principle of symmetric criticality [73] ensures that critical points of an invariant functional restricted to the space of equivariant trajectories are indeed free critical points. We stress that by the particular form of the interaction potentials, in our setting invariance is simply implied by equality of those masses which are interchanged by the action of  $G$  on the set of the indices. When the action of  $\sigma$  is transitive, all the masses must be equal and the associated  $G$ -equivariant trajectories give rise to *generalized choreographies*.

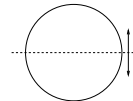
### Cyclic and Dihedral Actions

Consider the normal subgroup  $\ker \tau \triangleleft G$  and the quotient  $\bar{G} = G/\ker \tau$ . Since  $\bar{G}$  acts effectively on  $\mathbb{T}$ , it is either a *cyclic* group or a *dihedral* group.

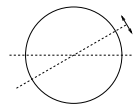
- If the group  $\bar{G}$  acts trivially on the orientation of  $\mathbb{T}$ , then  $\bar{G}$  is cyclic and we say that the action of  $G$  on  $\Lambda$  is of *cyclic type*.



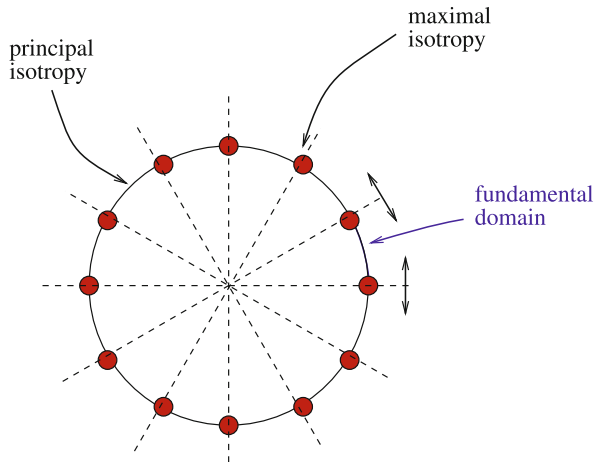
- If the group  $\bar{G}$  consists of a single reflection on  $\mathbb{T}$ , then we say that action of  $G$  on  $\Lambda$  is of *brake type*.



- Otherwise, we say that the action of  $G$  on  $\Lambda$  is of *dihedral type*.



With this first classification, we can easily associate to our minimization problem some proper boundary conditions.



**Proposition 1** Let  $\mathbb{I}$  be the fundamental domain (for a dihedral type), or any interval having as length the minimal angle of time-rotation in  $\bar{G}$  (for a cyclic type). Then the  $G$ -equivariant minimization problem is equivalent to problem of minimizing the action over all paths  $x: \mathbb{I} \rightarrow X^{\ker \tau}$  subject to the boundary conditions  $x(0) \in X^{H_0}$  and  $x(1) \in X^{H_1}$ , where  $H_0$  and  $H_1$  are the maximal isotropy subgroups of the boundary of  $\mathbb{I}$ .

### The Variational Approach

Let us consider the  $\mathcal{A}$  restricted to the space of symmetric loops  $\Lambda^G$ . We recall that the action functional is said

to be coercive if  $\lim_{|x| \rightarrow +\infty} \mathcal{A}(x) = +\infty$ . Coercivity implies the validity of the direct method of Calculus of Variations and, consequently, the existence of a minimizer.

**Proposition 2** *The action functional  $\mathcal{A}$  is coercive in  $\Lambda^G$  if and only if  $\chi^G = 0$ . Consequently, if  $\chi^G = 0$  then a minimizer of  $\mathcal{A}^G$  in  $\Lambda^G$  exists.*

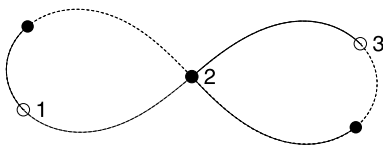
Given  $\rho$  and  $\sigma$ , we can compute  $\dim \chi^G$

$$\dim \chi^G = \frac{1}{|G|} \sum_{g \in G} \text{Tr}(\rho(g)) \# \text{Fix}(\sigma(g)) - d.$$

In the frame rotating with constant angular speed  $\omega$  the action  $\mathcal{A}^\omega$  is generally coercive, except for a (possible) discrete set values of  $\omega$ .

Of course, beside seeking minimizers, one can look for other type of critical points, such as mountain pass or others. As the potential  $U$  is singular on collisions, minimizers (or other critical points) can a priori be *collision trajectories*. Many strategies were proposed in the literature in order to overcome these obstacles. The development of a suitable Critical Point theory taking into account of the contribution of fake periodic solutions (the critical points at infinity) was proposed by some authors [9,60,77] and returned a good estimate of the number of periodic trajectories satisfying an appropriate bound on the length, with the main disadvantage of requiring a strong order of infinity at the collisions (the *strong force condition*). Another strategy to recover coercivity of action functional consists in imposing a symmetry constraint on the loop space. Surprisingly enough, once coercivity is recovered, also the problem of collisions becomes much less dramatic. This fact was remarked for the first time in [34] and widely exploited in the literature, also thanks to the neat idea, due to C. Marchal, of averaging over all possible variations (generalized and exposed in Sect. “Generalized Orbits and Singularities”) to avoid the occurrence of collisions for extremals of the action (Marchal idea was first exposed in [25]). This argument can be used in most of the known cases to prove that absence of collisions for minimizing trajectories and will be outlined in the last section.

### The Eight Shaped Three-Body Solution



In their paper [29], Chenciner and Montgomery exploited a variational argument in order to prove the existence of

a periodic trajectory for the three body problem, where the three particles move one a single eight-shaped curve, interchanging their positions after a fixed time. C. Moore [71] was the first to find numerically the eight, lead by topological reasons which turned out to be insufficient to insure its existence. One of the the simplest symmetries giving rise to the eight shaped trajectory is the following. Denote

$$x(t) = (x_1(t), x_2(t), x_3(t)) \in \mathbb{R}^6.$$

Let  $G = D_6$  be the dihedral group generated by two following reflections: the first

$$\begin{aligned} x_1(-t) &= -x_1(t), & x_2(-t) &= -x_3(t), \\ x_3(-t) &= -x_2(t), \end{aligned}$$

with  $g_1 = (\tau_1, \rho_1, \sigma_1)$  is  $\tau_1(t) = -t$ ,  $\rho_1(x) = -x$  and  $\sigma_1(1, 2, 3) = (1, 3, 2)$ . And the second

$$\begin{aligned} x_1(1-t) &= -x_2(t), & x_2(1-t) &= -x_1(t), \\ x_3(1-t) &= -x_3(t), \end{aligned}$$

with  $g_2 = (\tau_2, \rho_2, \sigma_2)$  is  $\tau_2(t) = 1-t$ ,  $\rho_2(x) = -x$  and  $\sigma_2(1, 2, 3) = (2, 1, 3)$ .

There are three possible groups yielding an eight-shaped trajectory. First, we consider the group of cyclic action type  $C_6$  (the *cyclic eight* having order 6), which acts cyclically on  $\mathbb{T}$  (i. e. by a rotation of angle  $\pi/3$ ), by a reflection in the plane  $E$ , and by the cyclic permutation  $(1, 2, 3)$  in the index set.

The second group, which we denote by  $D_{12}$ , is the group of order 12 obtained by extending  $C_6$  with the element  $h$  defined as follows:  $\tau(h)$  is a reflection in  $\mathbb{T}$ ,  $\rho(h)$  is the antipodal map in  $E$  (thus, the rotation of angle  $\pi$ ), and  $\sigma(h)$  is the permutation  $(1, 2)$ . This is the symmetry group used by Chenciner and Montgomery.

The third group is the subgroup of  $D_{12}$  generated by  $h$  and the subgroup  $C_3$  of order 3 of  $C_6 \subset D_{12}$ . We denote this group  $D_6$  (since it is a dihedral group of order 6). The symmetry groups  $D_{12}$  and  $D_6$  are of dihedral type. The choreography group  $C_3$  is a subgroup of all the three groups, thus the action is coercive on  $G$ -equivariant loops.

### The Rotating Circle Property (RCP)

As a second step of the variational approach to the  $n$ -body problem, one has to prove that the output of the minimization (or some other variational method) procedure is free of collisions. This point involves a deep analysis of the structure of the possible singularities and will be outlined in the last part of this article. After performing the analysis of collisions, we find that *if the action of  $G$  on  $\mathbb{T}$  and*

$X$  fulfills some conditions (computable) then (local) minimizers of the action functional  $\mathcal{A}$  in  $\Lambda^G \subset \Lambda$  do not have collisions.

For a group  $H$  acting orthogonally on  $\mathbb{R}^d$ , a circle  $\mathbb{S} \subset \mathbb{R}^d$  (with center in 0) is termed *rotating under  $H$*  if  $\mathbb{S}$  is invariant under  $H$  (that is, for every  $g \in H$   $g\mathbb{S} = \mathbb{S}$ ) and for every  $g \in H$  the restriction  $g|_{\mathbb{S}}: \mathbb{S} \rightarrow \mathbb{S}$  is a *rotation* (the identity is meant as a rotation of angle 0).

Let  $i \in \mathbf{n}$  be an index and  $H \subset G$  a subgroup. A circle  $\mathbb{S} \subset \mathbb{R}^d = V$  (with center in 0) is termed *rotating for  $i$  under  $H$*  if  $\mathbb{S}$  is rotating under  $H$  and

$$\mathbb{S} \subset V^{H_i} \subset V = \mathbb{R}^d,$$

where  $H_i \subset H$  denotes the *isotropy subgroup* of the index  $i$  in  $H$  relative to the action of  $H$  on the index set  $\mathbf{n}$  induced by restriction (that is, the isotropy  $H_i = \{g \in H | gi = i\}$ ).

**Definition 3** A group  $G$  acts with the *rotating circle property* if for every  $\mathbb{T}$ -isotropy subgroup  $G_i \subset G$  and for at least  $n - 1$  indexes  $i \in \mathbf{n}$  there exists in  $\mathbb{R}^d$  a rotating circle  $\mathbb{S}$  under  $G_i$  for  $i$ .

In most of the known examples the property is fulfilled. In [52] the following results were proved.

**Theorem 1** Consider a finite group  $K$  acting on  $\Lambda$  with the rotating circle property. Then a minimizer of the  $K$ -equivariant fixed-ends (Bolza) problem is free of collisions.

**Corollary 1** For every  $\alpha > 0$ , minimizers of the fixed-ends (Bolza) problem are free of interior collisions.

**Corollary 2** If the action of  $G$  on  $\Lambda$  is of cyclic type and  $\ker \tau$  has the rotating circle property then any local minimizer of  $\mathcal{A}^G$  in  $\Lambda^G$  is collisionless.

**Corollary 3** If the action of  $G$  on  $\Lambda$  is of cyclic type and  $\ker \tau = 1$  is trivial then any local minimizer of  $\mathcal{A}^G$  in  $\Lambda^G$  is collisionless.

**Theorem 2** Consider a finite group  $G$  acting on  $\Lambda$  so that every maximal  $\mathbb{T}$ -isotropy subgroup of  $G$  either has the rotating circle property or acts trivially on the index set  $\mathbf{n}$ . Then any local minimizer of  $\mathcal{A}^G$  yields a collision-free periodic solution of the Newton equations for the  $n$ -body problem in  $\mathbb{R}^d$ .

## Examples

In this section, for the sake of illustrating the power and the limitations of the approach through the  $G$ -equivariant minimization, we include a few examples fitting in our theoretical framework and we give some

hints of which symmetry groups satisfy or not assumptions of Theorem 2. Well-known examples are the celebrated Chenciner–Montgomery “eight” [29], Chenciner–Venturelli “Hip–Hop” solutions [30], Chenciner “generalized Hip–Hops” solutions [26], Chen’s orbit [20,22] and Terracini–Venturelli generalized Hip–Hops [85]. One word about the pictures of planar orbits: the configurations at the boundary points of the fundamental domain  $\mathbb{I}$  are denoted with an empty circle (starting point  $x_i(0)$ ) and a black disc (ending point  $x_i(t)$ , with  $t$  appropriate), with a label on the starting point describing the index of the particle. The trajectories of the particles with the times in  $\mathbb{I}$  are painted as thicker lines (thus it is possible to recover the direction of the movement from  $x_i(0)$  to  $x_i(t)$ ). Unfortunately this feature was not possible with the three-dimensional images.

Also, in all the following examples but 4 and 5 existence of the orbits follows directly from the results of Theorem 2. The existence of the orbits described in Examples 4 and 5, which goes beyond the scope of this article, has been recently proved by Chen in [22]. Thousands of other suitable actions and the corresponding orbits have been found by a special-purpose computer program based on GAP [53].

**Example 1 (Choreographies)** Consider the cyclic group  $G = \mathbb{Z}_n$  of order  $n$  acting trivially on  $V$ , with a cyclic permutation of order  $n$  on the index set  $\mathbf{n} = \{1, \dots, n\}$  and with a rotation of angle  $2\pi/n$  on the time circle  $\mathbb{T}$ . Since  $X^G = 0$ , by Proposition 2 the action functional  $\mathcal{A}^G$  is coercive. Moreover, since the action of  $G$  on  $\mathbb{T}$  is of cyclic type and  $\ker \tau = 1$ , by Corollary 2 the minimum exists and it has no collisions. For several numerical results and a description of choreographies we refer the reader to [32]. An insight of the variational properties of choreographies is provided in Sect. “Minimizing Properties of Simple Choreographies”.

**Example 2** Let  $n$  be odd. Consider the dihedral group  $G = D_{2n}$  of order  $2n$ , with the presentation

$$G = \langle g_1, g_2 | g_1^2 = g_2^n = (g_1 g_2)^2 = 1 \rangle.$$

Let  $\tau$  be the homomorphism defined by

$$\tau(g_1) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \text{ and } \tau(g_2) = \begin{bmatrix} \cos \frac{2\pi}{n} & -\sin \frac{2\pi}{n} \\ \sin \frac{2\pi}{n} & \cos \frac{2\pi}{n} \end{bmatrix}.$$

Furthermore, let the homomorphism  $\rho$  be defined by

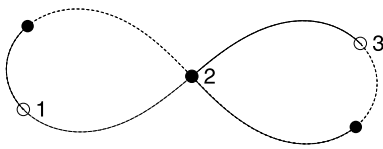
$$\rho(g_1) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } \rho(g_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Finally, let  $G$  act on  $\mathbf{n}$  by the homomorphism  $\sigma$  defined as  $\sigma(g_1) = (1, n-1)(2, n-2) \dots ((n-1)/2, (n+1)/2)$ ,

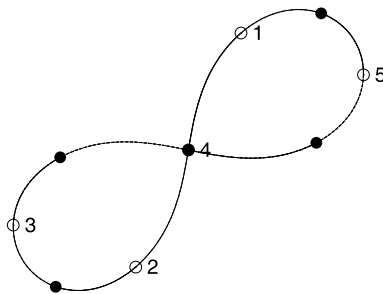
$\sigma(g_2) = (1, 2, \dots, n)$ , where  $(i_1, i_2, \dots, i_k)$  means the standard cycle-decomposition notation for permutation groups. By the action of  $g_2$  it is easy to show that all the loops in  $\Lambda^G$  are choreographies, and thus that, since  $\mathcal{X}^G = 0$ , the action functional is coercive. The maximal  $\mathbb{T}$ -isotropy subgroups are the subgroups of order 2 generated by the elements  $g_1 g_2^i$  with  $i = 0 \dots n-1$ . Since they are all conjugated, it is enough to show that one of them acts with the rotating circle property. Thus consider  $H = \langle g_1 \rangle \subset G$ . For every index  $i \in \{1, 2, \dots, n-1\}$  the isotropy  $H_i \subset H$  relative to the action of  $H$  on  $\mathbf{n}$  is trivial, and  $g_1$  acts by rotation on  $V = \mathbb{R}^2$ . Therefore for every  $i \in \{1, 2, \dots, n-1\}$  it is possible to choose a circle rotating under  $H$  for  $i$ , since, being  $H_i$  trivial,  $V^{H_i} = V$ . The resulting orbits are not homographic (since all the particles pass through the origin 0 at some time of the trajectory and the configurations are centered). For  $n = 3$  this is the *eight with less symmetry* of [25]. Possible trajectories are shown in Figs. 1 and 2.

**Example 3** As in the previous example, let  $n \geq 3$  be an odd integer. Let  $G = C_{2n} \cong \mathbb{Z}_2 + \mathbb{Z}_n$  be the cyclic group of order  $2n$ , presented as  $G = \langle g_1, g_2 | g_1^2 = g_2^n = g_1 g_2 g_1^{-1} g_2^{-1} = 1 \rangle$ . The action of  $G$  on  $\mathbb{T}$  is given by  $\tau(g_1 g_2) = \theta_{2n}$ , where  $\theta_{2n}$  denotes the rotation of angle  $\pi/n$  (hence the action will be of cyclic type). Now,  $G$  can act on the plane  $V = \mathbb{R}^2$  by the homomorphism  $\rho$  defined by

$$\rho(g_1) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \rho(g_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$



**n-Body Problem and Choreographies, Figure 1**  
The ( $D_6$ -symmetric) eight for  $n = 3$



**n-Body Problem and Choreographies, Figure 2**  
The ( $D_{10}$ -symmetric) eight with  $n = 5$

Finally, the action of  $G$  on  $\mathbf{n} = \{1, 2, \dots, n\}$  is given by the homomorphism  $\sigma: G \rightarrow \Sigma_n$  defined by  $\sigma(g_1) = ()$ ,  $\sigma(g_2) = (1, 2, \dots, n)$ . The cyclic subgroup  $H_2 = \langle g_2 \rangle \subset G$  gives the symmetry constraints of the choreographies, hence loops in  $\Lambda^G$  are choreographies and the functional is coercive. Furthermore, since the action is of cyclic type, by Corollary 3 the minimum of the action functional is collisionless. It is possible that such minima coincide with the minima of the previous example: this would imply that the symmetry group of the minimum contains the two groups above.

**Example 4** Consider four particles with equal masses and an odd integer  $q \geq 3$ . Let  $G = D_{4q} \times C_2$  be the direct product of the dihedral group of order  $4q$  with the group  $C_2$  of order 2. Let  $D_{4q}$  be presented by  $D_{4q} = \langle g_1, g_2 | g_1^2 = g_2^{2q} = (g_1 g_2)^2 = 1 \rangle$ , and let  $c \in C_2$  be the non-trivial element of  $C_2$ . Now define the homomorphisms  $\rho$ ,  $\tau$  and  $\sigma$  as follows:

$$\rho(g_1) = \tau(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

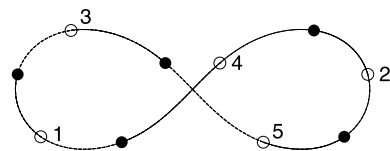
$$\rho(g_2) = \tau(g_2) = \begin{bmatrix} \cos \frac{2\pi}{2q} & -\sin \frac{2\pi}{2q} \\ \sin \frac{2\pi}{2q} & \cos \frac{2\pi}{2q} \end{bmatrix},$$

$$\rho(c) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \tau(c) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

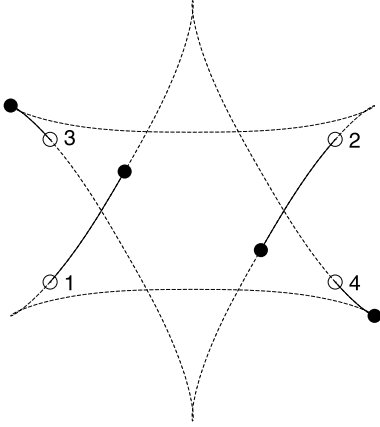
$$\sigma(g_1) = (1, 2)(3, 4), \quad \sigma(g_2) = (1, 3)(2, 4),$$

$$\sigma(c) = (1, 2)(3, 4).$$

It is not difficult to show that  $\mathcal{X}^G = 0$ , and thus the action is coercive. Moreover,  $\ker \tau = C_2$ , which acts on  $\mathbb{R}^2$  with the rotation of order 2, hence  $\ker \tau$  acts with the rotating circle property. Thus, by Proposition 2 and Theorem 1 the minimizer exists and does not have interior collisions. To exclude boundary collisions we cannot invoke Theorem 2, since the maximal  $\mathbb{T}$ -isotropy subgroups do not act with the rotating circle property. A possible graph for such a minimum can be found in Fig. 4, for  $q = 3$  (one needs to prove that the minimum is not the homographic solution – with a level estimate – and that there are no boundary collisions – with an argument similar to [20]).



**n-Body Problem and Choreographies, Figure 3**  
Another symmetry constraint for an eight-shaped orbit ( $n = 5$ )



**$n$ -Body Problem and Choreographies, Figure 4**  
The orbit of Example 4 with  $q = 3$

See also [22] for an updated and much generalized treatment of such orbits.

**Example 5** Consider four particles with equal masses and an even integer  $q \geq 4$ . Let  $G = D_q \times C_2$  be the direct product of the dihedral group of order  $2q$  with the group  $C_2$  of order 2. Let  $D_{4q}$  be presented by  $D_{4q} = \langle g_1, g_2 | g_1^2 = g_2^q = (g_1 g_2)^2 = 1 \rangle$ , and let  $c \in C_2$  be the non-trivial element of  $C_2$ . As in Example 4, define the homomorphisms  $\rho, \tau$  and  $\sigma$  as follows.

$$\rho(g_1) = \tau(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

$$\rho(g_2) = \tau(g_2) = \begin{bmatrix} \cos \frac{2\pi}{q} & -\sin \frac{2\pi}{q} \\ \sin \frac{2\pi}{q} & \cos \frac{2\pi}{q} \end{bmatrix},$$

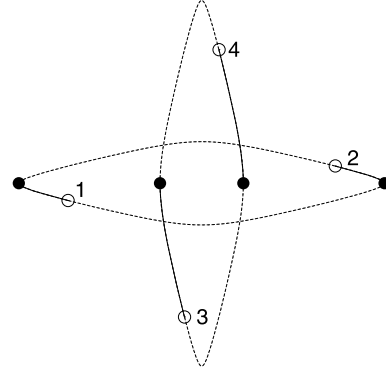
$$\rho(c) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \tau(c) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\sigma(g_1) = (1, 2)(3, 4), \quad \sigma(g_2) = (1, 3)(2, 4),$$

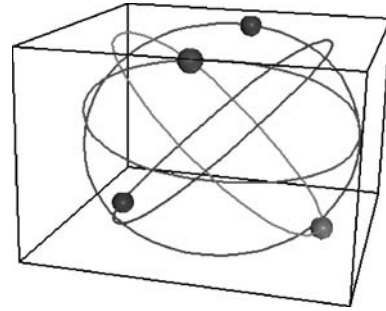
$$\sigma(c) = (1, 2)(3, 4).$$

Again, one can show that a minimizer without interior collisions exists since  $\ker \tau = C_2$  acts with the rotating circle property (a possible minimizer is shown in Fig. 5). This generalizes Chen's orbit [22]. See also [20].

**Example 6 (Hip-Hops)** If  $G = \mathbb{Z}_2$  is the group of order 2 acting trivially on  $\mathbf{n}$ , acting with the antipodal map on  $V = \mathbb{R}^3$  and on the time circle  $\mathbb{T}$ , then again  $X^G = 0$ , so that Proposition 2 holds. Furthermore, since the action is of cyclic type Corollary 2 ensures that minimizers have no collisions. Such minimizers were called *generalized Hip-Hops* in [25]. See also [26]. A subclass of symmetric trajectories leads to a generalization of such a Hip-



**$n$ -Body Problem and Choreographies, Figure 5**  
A possible minimizer for Example 5



**$n$ -Body Problem and Choreographies, Figure 6**  
The Chenciner-Venturelli Hip-Hop

Hop. Let  $n \geq 4$  an even integer. Consider  $n$  particles with equal masses, and the group  $G = C_n \times C_2$  direct product of the cyclic group of order  $n$  (with generator  $g_1$ ) and the group  $C_2$  of order 2 (with generator  $g_2$ ). Let the homomorphisms  $\rho, \sigma$  and  $\tau$  be defined by

$$\rho(g_1) = \begin{bmatrix} \cos \frac{2\pi}{n} & -\sin \frac{2\pi}{n} & 0 \\ \sin \frac{2\pi}{n} & \cos \frac{2\pi}{n} & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\rho(g_2) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad \tau(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\tau(g_2) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$

$$\sigma(g_1) = (1, 2, 3, 4), \quad \sigma(g_2) = ().$$

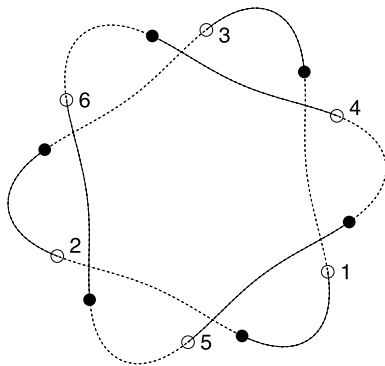
It is easy to see that  $X^G = 0$ , and thus a minimizer exists. Since the action is of cyclic type, it suffices to exclude interior collisions. But this follows from the fact that  $\ker \tau = C_n$  has the rotating circle property. This example is the natural generalization of the Hip-Hop solution of [30] to  $n \geq 4$  bodies. We can see the trajectories in Fig. 6.



**Example 7** Consider the direct product  $G = D_6 \times C_3$  of the dihedral group  $D_6$  (with generators  $g_1$  and  $g_2$  of order 3 and 2 respectively) of order 6 and the cyclic group  $C_3$  of order 3 generated by  $c \in C_3$ . Let us consider the planar  $n$ -body problem with  $n = 6$  with the symmetry constraints given by the following  $G$ -action.

$$\begin{aligned}\rho(g_1) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & \rho(g_2) &= \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \\ \rho(c) &= \begin{bmatrix} \cos \frac{2\pi}{3} & -\sin \frac{2\pi}{3} \\ \sin \frac{2\pi}{3} & \cos \frac{2\pi}{3} \end{bmatrix}, \\ \tau(g_1) &= \begin{bmatrix} \cos \frac{2\pi}{3} & -\sin \frac{2\pi}{3} \\ \sin \frac{2\pi}{3} & \cos \frac{2\pi}{3} \end{bmatrix}, \\ \tau(g_2) &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, & \tau(c) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \\ \sigma(g_1) &= (1, 3, 2)(4, 5, 6), & \sigma(g_2) &= (1, 4)(2, 5)(3, 6), \\ \sigma(c) &= (1, 2, 3)(4, 5, 6).\end{aligned}$$

By Proposition 2 one can prove that a minimizer exists, and since  $G$  acts with the rotating circle property (actually, the elements of the image of  $\rho$  are rotations) on  $\mathbb{T}$ -maximal isotropy subgroups, the conclusion of Theorem 2 holds. It is not difficult to see that configurations in  $\mathcal{X}^{\ker \tau}$  are given by two centered equilateral triangles. Now, to guarantee that the minimizer is not a homographic solution, of course it suffices to show that there are no homographic solutions in  $\Lambda^G$  (like in the case of Example 2). This follows from the easy observation that at some times  $t \in \mathbb{T}$  with maximal isotropy it happens that  $x_1 = -x_4$ ,  $x_2 = -x_5$  and  $x_3 = -x_6$ , while at some other times it happens that  $x_1 = -x_5$ ,  $x_2 = -x_6$  and  $x_3 = -x_4$  or that  $x_1 = -x_6$ ,  $x_2 = -x_4$  and  $x_3 = -x_5$  and this implies that there are no homographic loops in  $\Lambda^G$ . With no difficulties the same action can be defined for  $n = 2k$ , where  $k$  is



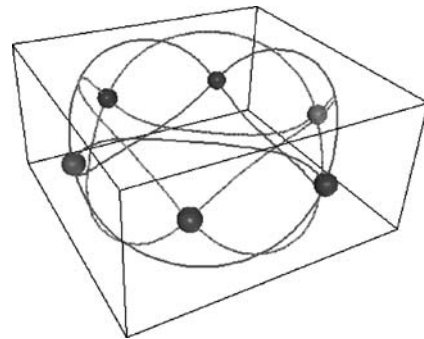
**n-Body Problem and Choreographies, Figure 7**  
The planar equivariant minimizer of Example 7

any odd integer. We can see a possible trajectory in Fig. 7. Also, it is not difficult to consider a similar example in dimension 3. With  $n = 6$  and the notation of  $D_6$  and  $C_3$  as above, consider the group  $G = D_6 \times C_3 \times C_2$ . Let  $g_1, g_2, c$  be as above, and let  $c_2$  be the generator of  $C_2$ . The homomorphisms  $\rho, \tau$  and  $\sigma$  are defined in a similar way by

$$\begin{aligned}\rho(g_1) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, & \rho(g_2) &= \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\ \rho(c) &= \begin{bmatrix} \cos \frac{2\pi}{3} & -\sin \frac{2\pi}{3} & 0 \\ \sin \frac{2\pi}{3} & \cos \frac{2\pi}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\ \rho(c_2) &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \\ \tau(g_1) &= \begin{bmatrix} \cos \frac{2\pi}{3} & -\sin \frac{2\pi}{3} \\ \sin \frac{2\pi}{3} & \cos \frac{2\pi}{3} \end{bmatrix}, & \tau(g_2) &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \\ \tau(c) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & \tau(c_2) &= \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \\ \sigma(g_1) &= (1, 3, 2)(4, 5, 6), & \sigma(g_2) &= (1, 4)(2, 5)(3, 6), \\ \sigma(c) &= (1, 2, 3)(4, 5, 6), & \sigma(c_2) &= ().\end{aligned}$$

In the resulting collisionless minimizer (again, it follows by Proposition 2 and Theorem 2) two equilateral triangles rotate in opposite directions and have a “brake” motion on the third axis. The likely shape of the trajectories can be found in Fig. 8.

**Example 8 (Marchal’s  $P_{12}$ -symmetry revisited)** Let  $k \geq 2$  be an integer, and consider the cyclic group  $G = C_{6k}$  of order  $6k$  generated by the element  $c \in G$ . Now consider orbits for  $n = 3$  bodies in the space of dimension  $d = 3$ . With a minimal effort and suitable changes the example can be generalized for every  $n \geq 3$ . We leave the details to



**n-Body Problem and Choreographies, Figure 8**  
The three-dimensional equivariant minimizer of Example 7

the reader. The homomorphisms  $\rho$ ,  $\tau$  and  $\sigma$  are defined by

$$\rho(c) = \begin{bmatrix} \cos \frac{\pi}{k} & -\sin \frac{\pi}{k} & 0 \\ \sin \frac{\pi}{k} & \cos \frac{\pi}{k} & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

$$\tau(c) = \begin{bmatrix} \cos \frac{2\pi}{6k} & -\sin \frac{2\pi}{6k} \\ \sin \frac{2\pi}{6k} & \cos \frac{2\pi}{6k} \end{bmatrix},$$

$$\sigma(c) = (1, 2, 3).$$

Straightforward calculations show that  $\mathcal{X}^G = 0$  and hence Proposition 2 can be applied. Furthermore, the action is of cyclic type with  $\ker \tau = 1$ , and hence by Corollary 2 the minimizer does not have collisions. It is left to show that this minimum is not a homographic motion. The only homographic motion in  $\Lambda^G$  is a Lagrange triangle  $y(t) = (y_1, y_2, y_3)(t)$ , rotating with angular velocity  $3 - 2k$  (assume that the period is  $2\pi$ , i. e. that  $T = |\mathbb{T}| = 2\pi$ ) in the plane  $u_3 = 0$  (let  $u_1, u_2, u_3$  denote the coordinates in  $\mathbb{R}^3$ ). To be a minimum it needs to be inscribed in the horizontal circle of radius  $((\alpha 3^{-\alpha/2})/(2(3-2k)^2))^{1/(2+\alpha)}$ . Now, for every function  $\phi(t)$  defined on  $\mathbb{T}$  such that  $\phi(c^3 t) = -\phi(t)$ , the loop given by  $v_1(t) = (0, 0, \phi(t))$ ,  $v_2(t) = (0, 0, \phi(c^{-2}t))$  and  $v_3(t) = (0, 0, \phi(c^2 t))$  is  $G$ -equivariant, and thus belongs to  $\Lambda^G$ . If one computes the value of Hessian of the Lagrangian action  $\mathcal{A}$  in  $y$  and in the direction of the loop  $v$  one finds that

$$D_v^2 \mathcal{A}|_y = 3 \int_0^{2\pi} \dot{\phi}^2(t) dt - 2(3-2k)^2 \times \int_0^{2\pi} (\phi(t) + \phi(ct))^2 dt.$$

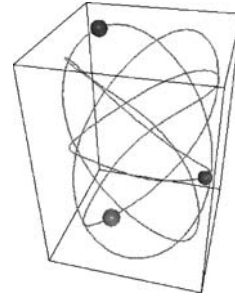
In particular, if we set the function  $\phi(t) = \sin(kt)$ , which has the desired property, elementary integration yields

$$D_v^2 \mathcal{A}|_y = 3\pi(k^2 - 2(3-2k)^2),$$

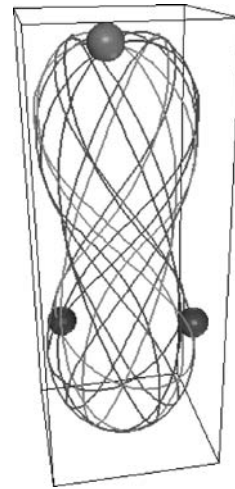
which does not depend on  $\alpha$  and is negative for every  $k \geq 3$ . Thus for every  $k \geq 3$  the minimizer is not homographic. We see a possible trajectory in Fig. 9.

**Remark 1** In the previous example, if  $k \not\equiv 0 \pmod{3}$ , the cyclic group  $G$  can be written as the sum  $C_3 + C_{2k}$ . The generator of  $C_3$  acts trivially on  $V$ , acts with a rotation of order 3 on  $\mathbb{T}$  and with the cyclic permutation  $(1, 2, 3)$  on  $\{1, 2, 3\}$ . This means that for all  $k \not\equiv 0 \pmod{3}$  the orbits of Example 8 are non-planar choreographies. Furthermore, it is possible to define a cyclic action of the same kind by setting  $\tau$  and  $\sigma$  as above and

$$\rho(c) = \begin{bmatrix} \cos \frac{p}{3k} \pi & -\sin \frac{p}{3k} \pi & 0 \\ \sin \frac{p}{3k} \pi & \cos \frac{p}{3k} \pi & 0 \\ 0 & 0 & -1 \end{bmatrix},$$



**$n$ -Body Problem and Choreographies, Figure 9**  
The non-planar choreography of Example 8 for  $k = 4$



**$n$ -Body Problem and Choreographies, Figure 10**  
A non-planar symmetric orbit of Remark 1, with  $k = 3$  and  $p = 1$

where  $p$  is a non-zero integer. If  $p = 3$  one obtains the same action as in Example 8. One can perform similar computations and obtain that the Lagrange orbit (with angular velocity  $p - 2k$ , this time) is not a minimizer for all  $(p, k)$  such that  $0 < p < 3k$  and  $k^2 - 2(p - 2k)^2 < 0$ .

### More Examples with non Trivial Core

Now consider the *core* of  $G$ ,  $\ker \tau = K \subset G$ . With an abuse of notation, using the factorization of the action, we can identify  $K$  with a subgroup of  $O(3)$ . A very remarkable class of symmetry groups with non trivial core has been defined by Ferrario in [50], through its *Krh* decomposition:

**Definition 4** Let  $G$  be a symmetry group. Then define

1.  $K = \ker \tau$ ,
2.  $[r] \in W_{O(3)} K$  as the image in the Weyl group of the generator mod  $K$  of  $\ker \det \tau \subset G/K$  (corresponding

to the time-shift with minimal angle). If  $\ker \det \tau = K$ , then  $[r] = 1$ .

3.  $[h] \in W_{O(3)}K$  as the image in the Weyl group of one of the time-reflections mod  $K$  in  $G/K$ , in the cases such an element exists. Otherwise it is not defined.

In short, the triple

$$(K, [r], [h])$$

is said the  $Krh$  data of  $G$ .

**Example 9**

Consider the icosahedral group  $Y$  of order 60. The group  $G$  with  $Krh$  data

$$\begin{pmatrix} 1 & (1, -1) \\ Y & -1 \end{pmatrix}$$

is isomorphic to the direct product  $I \times Y$  of order 120, and acts on the euclidean space  $\mathbb{R}^3$  as the full icosahedron group. The action on  $\mathbb{T}$  is cyclic and given by the fact that  $\ker \tau = 1 \times Y$ . The isotropy is generated by the central inversion  $-1$ , and hence the set of bodies is  $G/I \cong Y$ . Thus at any time  $t$  the 60 point particles are constrained to be a  $Y$ -orbit in  $\mathbb{R}^3$  (which does not mean they are vertices of a icosahedron, simply that the configuration is  $Y$ -equivariant). After half period every body is in the antipodal position:  $x_i(t + T/2) = -x_i$  (in other words, the group contains the *anti-symmetry*), also known as *Italian symmetry* – see [25,26]. Of course, the group  $Y$  is just an example: one can choose also the tetrahedral group  $T$  or the octahedral  $O$  and obtain anti-symmetric orbits for 12 (tetrahedral) or 24 (octahedral) bodies, as depicted in Fig. 11. The action is by its definition transitive and coercive; local minimizers are collisionless since the maximal

$\mathbb{T}$ -isotropy group acts as a subgroup of  $SO(3)$  (i. e. orientation-preserving).

**Example 10** Let  $G$  be the group with  $Krh$  data

$$\begin{pmatrix} 1 & (1, -1) \\ D_k & -1 \end{pmatrix},$$

where  $D_k$  is the rotation dihedral group of order  $2k$ . As in the previous example, the action is such that the action functional is coercive and its local minima collisionless. At every time instant the bodies are  $D_k$ -equivariant in  $\mathbb{R}^k$  and the anti-symmetry holds. Approximations of minima can be seen in Fig. 12.

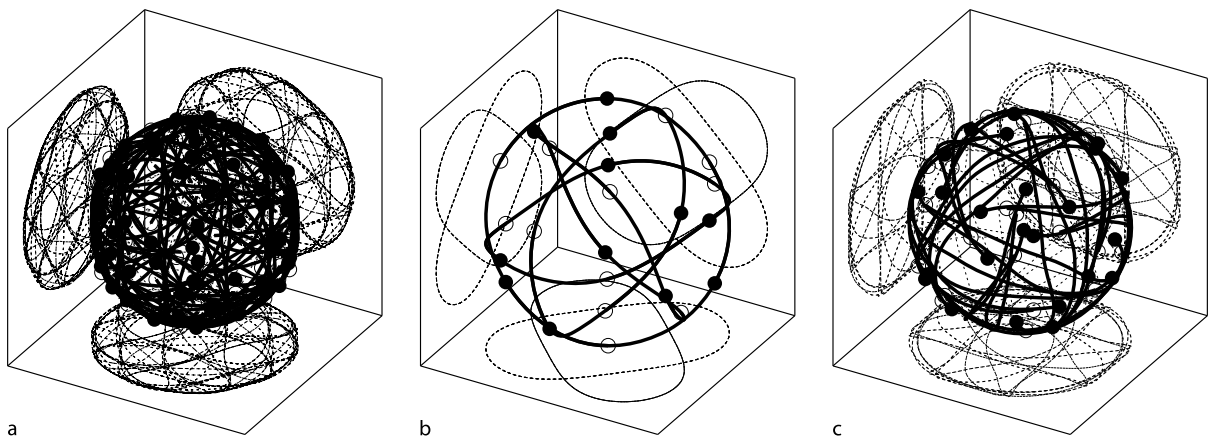
**Example 11** To illustrate the case of non-transitive symmetry group, consider the following (cyclic)  $Krh$  data:

$$\begin{pmatrix} 1 & (3, -1) \\ 1 & -1 \end{pmatrix},$$

which yields a group of order 6 acting cyclically on 3 bodies, and with the antipodal map on  $\mathbb{R}^3$ . Since  $\ker \tau$  is trivial and the group is of cyclic type, local minima are collisionless. Now, by adding  $k$  copies of such group one obtains a symmetry group having  $k$  copies of it as its transitive components, where still local minimizers are collisionless and the restricted functional is coercive. Some possible minima can be found in Fig. 13, for  $k = 3, 4$ .

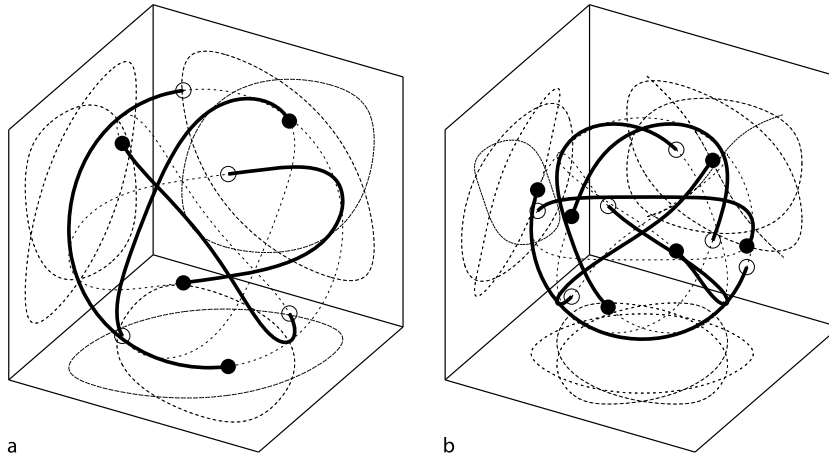
### The 3-Body Problem

The major achievement of [14] is to give the complete description of the outcome of the equivariant minimization procedure for the planar three-body problem. First we can ensure that minimizers are always collisionless.

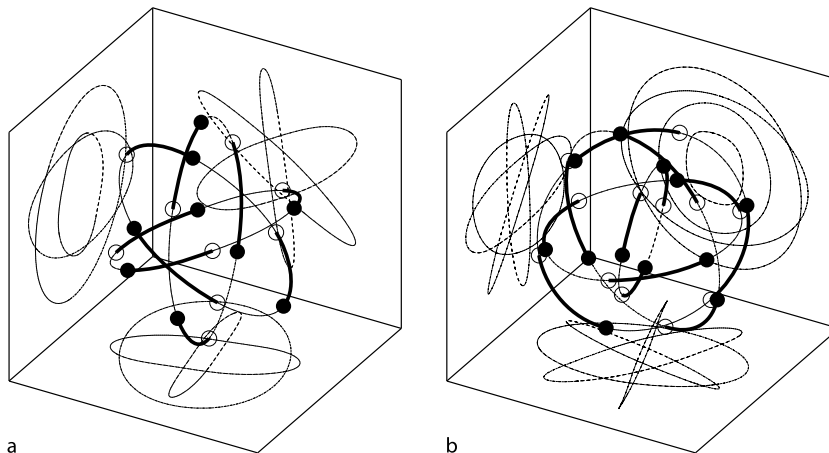


**n-Body Problem and Choreographies, Figure 11**

60-icosahedral  $Y$ , 12-tetrahedral  $T$  and 24-octahedral  $O$  periodic minimizers (chiral)



$n$ -Body Problem and Choreographies, Figure 12  
4-dihedral  $D_2$  and 6-dihedral  $D_6$  symmetric periodic minimizers



$n$ -Body Problem and Choreographies, Figure 13  
9 and 12 bodies in anti-choreographic constraints grouped by 3

**Theorem 3** Let  $G$  a symmetry group of the Lagrangian in the 3-body problem (in a rotating frame or not). If  $G$  is not bound to collision (i. e. every equivariant loop has collisions), then any (possible) local minimizer is collisionless.

A symmetry group  $G$  of the Lagrangian functional  $\mathcal{A}$  is termed

- *bound to collisions* if all  $G$ -equivariant loops actually have collisions,
- *fully uncoercive* if for every possible rotation vector  $\omega$  the action functional  $\mathcal{A}_\omega^G$  in the frame rotating around  $\omega$  with angular speed  $|\omega|$  is not coercive in the space of  $G$ -equivariant loops (that is, its global minimum escapes to infinity);
- *homographic* if all  $G$ -equivariant loops are constant up to orthogonal motions and rescaling.

- The *core* of the group  $G$  is the subgroup of all the elements which do not move the time  $t \in \mathbb{T}$ .

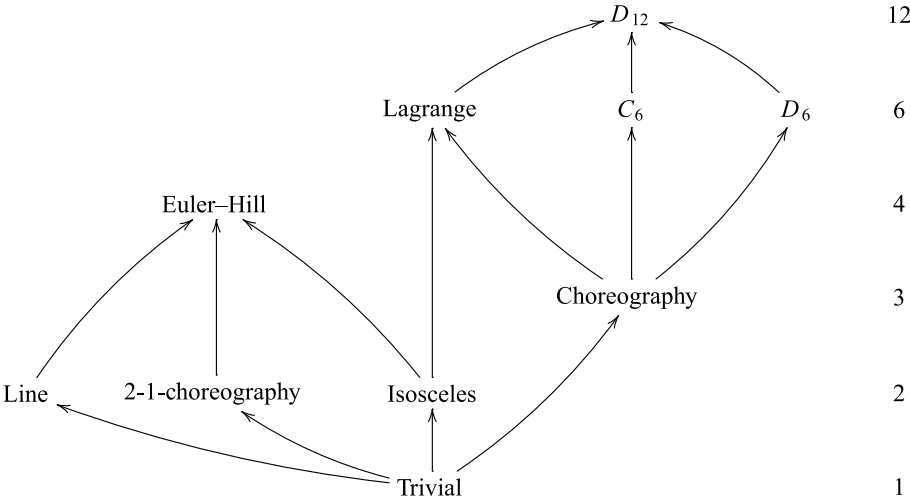
If, for every angular velocity,  $G$  is a symmetry group for the Lagrangian functional in the rotating frame, then we will say that  $G$  is of type  $R$ . This is a fundamental property for symmetry groups. In fact, if  $G$  is *not* of type  $R$ , it turns out that the angular momentum of all  $G$ -equivariant trajectories vanishes.

### The Classification of Planar Symmetry Groups for 3-body

**Theorem 4** Let  $G$  be a symmetry group of the Lagrangian action functional in the planar 3-body problem. Then, up to a change of rotating frame,  $G$  is either bound to collisions, fully uncoercive, homographic, or conjugated to one of the

*n*-Body Problem and Choreographies, Table 1  
Planar symmetry groups with trivial core

Name	G	Type <i>R</i>	Action type	trans. dec.	RCP	HGM
Trivial	1	yes		1 + 1 + 1	yes	yes
Line	2	yes	brake	1 + 1 + 1	(no)	no
2 – 1-choreography	2	yes	cyclic	2 + 1	yes	no
Isosceles	2	yes	brake	2 + 1	no	yes
Hill	4	yes	dihedral	2 + 1	no	no
3-choreography	3	yes	cyclic	3	yes	yes
Lagrange	6	yes	dihedral	3	no	yes
$C_6$	6	no	cyclic	3	yes	no
$D_6$	6	no	dihedral	3	yes	no
$D_{12}$	12	no	dihedral	3	no	no



*n*-Body Problem and Choreographies, Figure 14  
The poset of symmetry groups for the planar 3-body problem

symmetry groups listed in Table 1 (RCS stands for Rotating Circle Property and HGM for Homographic Global Minimizer).

Planar Symmetry Groups

- *The trivial symmetry*  
Let  $G$  be the trivial subgroup of order 1. It is clear that it is of type  $R$ , it has the rotating circle property. It yields a coercive functional on  $\Lambda^G = \Lambda$  only when  $\omega$  is not an integer. If  $\omega = \frac{1}{2} \pmod{1}$  then the minimizers are minimizers for the anti-symmetric symmetry group (also known as *Italian symmetry*)  $x(at) = ax(t)$ , where  $a$  is the antipodal map on  $\mathbb{T}$  and  $E$ . The masses can be different.

**Proposition 3** For every  $\omega \notin \mathbb{Z}$  and every choice of masses the minimum for the trivial symmetry occurs in the relative

equilibrium motion associated to the Lagrange central configuration.

- *The line symmetry*  
Another case of symmetry group that can be extended to rotating frames with arbitrary masses is the line symmetry: the group is a group of order 2 acting by a reflection on the time circle  $\mathbb{T}$ , by a reflection on the plane  $E$ , and trivially on the set of indexes. That means, at time 0 and  $\pi$  the masses are collinear, on a fixed line  $l \subset E$ . It is coercive only when  $\omega \notin \mathbb{Z}$ . In this case the Lagrangian solution cannot be a minimum, while it can be the relative equilibrium associated with the Euler configuration.
- *The 2 – 1-choreography symmetry*  
Consider the group of order 2 acting as follows:  $\rho(g) = 1$ ,  $\tau(g) = -1$  (that is, the translation of half-period) and  $\sigma(g) = (1, 2)$  (that is,  $\sigma(g)(1) = 2$ ,  $\sigma(g)(2) = 1$ ,



and  $\sigma(g)(3) = 3$ ). That is, it is a half-period choreography for the bodies 1 and 2. It can be extended to rotating frames and coercive for a suitable choice of  $\omega \not\equiv 0, 1 \pmod 2$ .

The Euler's orbit with  $k = 1$  and the Hill's orbits with  $k = \pm 1$  are equivariant for the 2 – 1-choreography symmetry, while the Euler's orbit with  $k = 0$  is not equivariant for this symmetry. In Figures 15 and 16 Euler 1 represents the action levels on the Euler's orbit with  $k = 1$ , Hill 1–2 the ones on the Hill's with  $k = \pm 1$ .

- *The isosceles symmetry*

The isosceles symmetry can be obtained as follows: the group is of order 2, generated by  $h$ ;  $\tau(h)$  is a reflection in the time circle  $\mathbb{T}$ ,  $\rho(h)$  is a reflection along a line  $l$  in  $E$ , and  $\sigma(h) = (1, 2)$  as above. The constraint is therefore that at time 0 and  $\pi$  the 3-body configuration

is an isosceles triangle with one vertex on 1 (the third).

**Proposition 4** *For every  $\omega \notin \mathbb{Z}$  and every choice of masses the minimum for the isosceles symmetry occurs in the relative equilibrium motion associated to the Lagrange configuration.*

- *The Euler–Hill symmetry*

Now consider the symmetry group with a cyclic generator  $r$  of order 2 (i. e.  $\tau(r) = -1$ ) and a time reflection  $h$  (i. e.  $\tau(h)$  is a reflection of  $\mathbb{T}$ ) given by  $\rho(r) = 1$ ,  $\sigma(r) = (1, 2)$ ,  $\rho(h)$  is a reflection and  $\sigma(h) = ()$ . It contains the 2 – 1-choreography (as the subgroup  $\ker \det(\tau)$ ), the isosceles symmetry (as the isotropy of  $\pi/2 \in \mathbb{T}$ ) and the line symmetry (as the isotropy of  $0 \in \mathbb{T}$ ) as subgroups.

**Proposition 5** *The minimum of the Euler–Hill symmetry is not homographic, provided that the angular velocity  $\omega$  is close to 0.5 and the values of the masses are close to 1.*

- *The choreography symmetry*

The choreography symmetry is given by the group  $C_3$  of order 3 acting trivially on the plane  $E$ , by a rotation of order 3 in the time circle  $\mathbb{T}$  and by the cyclic permutation  $(1, 2, 3)$  of the indices.

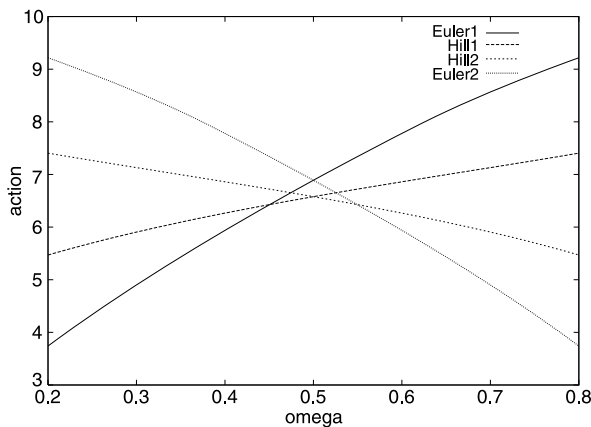
**Proposition 6** *For every  $\omega$  the minimal choreography of the 3 body problem is a rotating Lagrange configuration.*

- *The Lagrange symmetry*

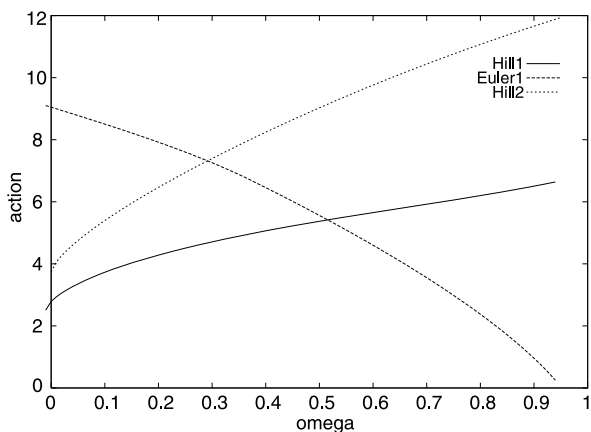
The Lagrange symmetry group is the extension of the choreography symmetry group by the isosceles symmetry group. Thus, it is a dihedral group of order 6, the action is of type R. Hence, the relative equilibrium motions associated to the Lagrange configuration are admissible motions for this symmetry and, again, the minimizer occurs in the relative equilibrium motion associated to the Lagrange configuration.

- *The Chenciner–Montgomery symmetry group and the eights*

There are three symmetry groups (up to change of coordinates) that yield the Chenciner–Montgomery figure eight orbit: they are the only symmetry groups which do not extend to the rotating frame and we have already described them in Subsect. “[The Eight Shaped Three-Body Solution](#)”. One can prove that all G-equivariant trajectories have vanishing angular momentum, whenever the group is not of type R. Moreover, we were able to partially answer to the open question (posed by



*n*-Body Problem and Choreographies, Figure 15  
Action levels for the line symmetry



*n*-Body Problem and Choreographies, Figure 16  
Action levels for the 2 – 1-choreography symmetry

Chenciner) whether their minimizers coincide or not: for two of them ( $D_6$  and  $D_{12}$ ) the minimizer is necessarily the same.

### Space Three-body Problem

Based on the classification of planar groups, by introducing a natural notion of space extension of a planar group, Ferrario gave in [51] a complete answer to the classification problem for the three-body problem in the space and at the same time to determine the resulting minimizers and describe its more relevant properties.

**Theorem 5** *Symmetry groups not bound to collisions, not fully uncoercive and not homographic are, up to a change of rotating frame, either the three-dimensional extensions of planar groups (if trivial core) listed in table below or the vertical isosceles triangle (if non-trivial core).*

The next theorem is the answer to the natural questions about collisions and description of some main features of minimizers.

**Theorem 6** *Let  $G$  be a symmetry group not bound to collisions and not fully uncoercive. Then*

1. *Local minima of  $\mathcal{A}^\omega$  do not have collisions.*
2. *In the following cases minimizers are planar trajectories:*
  - a) *If  $G$  is not of type R:  $D_6^{+, -}$ ,  $D_6^{-, +}$  and  $D_{12}^{-, +}$  (and then  $G$ -equivariant minimizers are Chenciner–Montgomery eights).*
  - b) *If there is a  $G$ -equivariant minimal Lagrange rotating solution:  $C_1^-$ ,  $H_2^{+, -}$ ,  $C_3^+$ ,  $L_6^{+, +}$ ,  $L_6^{+, -}$  (and then the Lagrange solution is of course the minimizer).*
  - c) *If the core is non-trivial and it is not the vertical isosceles (and then minimizers are homographic).*
3. *In the following cases minimizers are always non-planar:*
  - a) *The groups  $L_6^{-, +}$  and  $C_3^-$  for all  $\omega \in (-1, 1) + 6\mathbb{Z}$ ,  $\omega \neq 0$  (the minimizers for  $L_6^{-, +}$  are the elements of*

*Marchal family  $P_{12}$ , and minimizers of  $C_3^-$  are a less-symmetric family  $P_{12}'$ ).*

- b) *The extensions of line and Hill–Euler type groups, for an open subset of mass distributions and angular speeds  $\omega$ :  $L_2^{+, -}$ ,  $L_2^{-, +}$ ,  $H_4^{+, -}$  and  $H_4^{-, +}$  (for  $L_2^{-, +}$  this happens also with equal masses).*
- c) *The vertical isosceles for suitable choices of masses and  $\omega$ .*

### Minimizing Properties of Simple Choreographies

In the space of symmetric (choreographical) loops, the action takes the form

$$\mathcal{A}(x) = \frac{1}{2} \sum_{h=0}^{n-1} \int_0^\tau |\dot{x}(t + h\tau)|^2 dt + \frac{1}{2} \times \sum_{\substack{h, l=0 \\ h \neq l}}^{n-1} \int_0^\tau \frac{dt}{|x(t + l\tau) - x(t + h\tau)|^\alpha}.$$

A natural question concerns the nature of the minimizers under the sole choreographical constraint. Unfortunately, the bare minimization among choreographical loops gives returns only trivial motions:

**Theorem 7** *For every  $\alpha \in \mathbb{R}_+^*$  and  $d \geq 2$ , the absolute minimum of  $\mathcal{A}$  on  $\Lambda$  is attained on a relative equilibrium motion associated to the regular  $n$ -gon.*

This theorem extends some related result for the Italian symmetry by Chenciner and Desolneux [27], and the results in Sect. “The 3-Body Problem”. The proof is based on a (quite involved) convexity argument together with the analysis of some spectral properties related to the choreographical constraint. Now, in order to find nontrivial minimizers, we look at the same problem in a rotating frame. In order to take into account of the Coriolis force, the new action functional has to contain a gyroscopic term:

$$\mathcal{A}(y) = \frac{1}{2} \int_0^{2\pi} |\dot{y}(t) + J\omega y(t)|^2 dt + \frac{1}{2} \sum_{h=1}^{n-1} \int_0^{2\pi} \frac{dt}{|y(t) - y(t + h\tau)|^\alpha}.$$

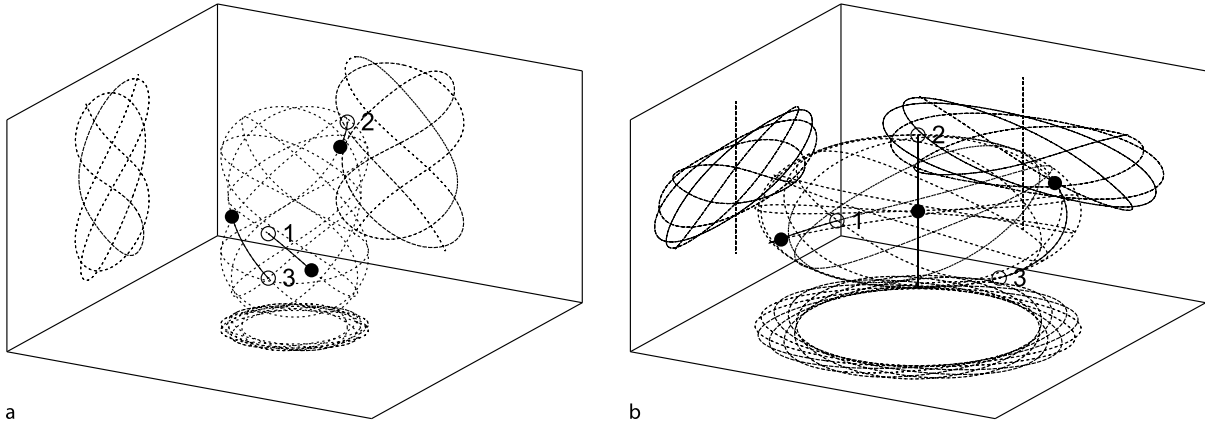
Consider the function  $h: \mathbb{R}_+^* \rightarrow \mathbb{N}$ ,  $h(\omega) = \min_{n \in \mathbb{N}^*} (\omega - n)^2/n^2$  and let  $\omega^* = \frac{4}{3}$ . The same technique used

<sup>2</sup>Highly likely they are not distinct families: this is the recurring phenomenon of “more symmetries than expected” in  $n$ -body problems.

### n-Body Problem and Choreographies, Table 2

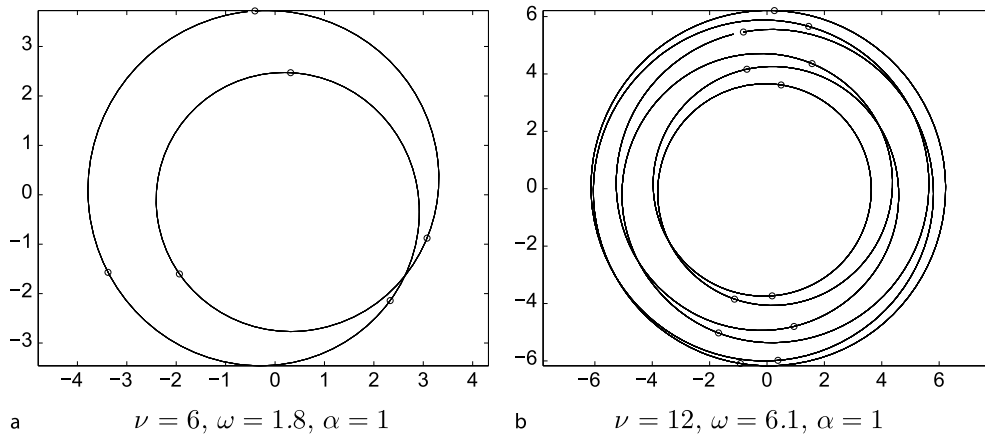
Space extensions of planar symmetry groups with trivial core

Name	Extensions
Trivial	$C_1^-$
Line	$L_2^{+, -}, L_2^{-, +}$
Isosceles	$H_2^{+, -}, H_2^{-, +}$
Hill	$H_4^{+, -}, H_4^{-, +}$
3-choreography	$C_3^+, C_3^-$
Lagrange	$L_6^{+, +}, L_6^{+, -}, L_6^{-, +}$
$D_6$	$D_6^{+, -}, D_6^{-, +}$
$D_{12}$	$D_{12}^{-, +}$



$n$ -Body Problem and Choreographies, Figure 17

Non planar minimizers for the groups  $L_6^{-,+}$  and  $L_2^{+,-}$ , plotted in the fixed frame



$n$ -Body Problem and Choreographies, Figure 18

Examples for Theorem 10 close to an integer that divides  $n$

for the inertial system extends to rotating systems having small angular velocity; this gives the following result.

**Theorem 8** *If  $\omega \in (0, \omega^*) \setminus \{1\}$ , then the action attains its minimum on a circle with minimal period  $2\pi$  and radius depending on  $n$ ,  $\alpha$  and  $\omega$ .*

### When $\omega$ is Close to an Integer

The situation changes dramatically when  $\omega$  is close to some integer. To understand this phenomenon, let us first check the result of the minimization procedure when  $\omega$  is an integer:

#### Proposition 7

(1) *If  $\omega = n$ , then the action has a continuum of minimizers.*

- (2) *If  $\omega = k$ , coprime with  $n$ , then the action does not achieve its infimum (escape of minimizing sequences).*
- (3) *If  $\omega = k$  and  $k$  divides  $n$ , then the action does not achieve its infimum (clustering of minimizing sequences).*

As a consequence, we have the following result:

**Theorem 9** *Suppose that  $n$  and  $k$  are coprime. Then there exist  $\epsilon = \epsilon(\alpha, n, k)$  such that if  $\omega \in (k - \epsilon, k + \epsilon)$  the minimum of the action is attained on a circle with minimal period  $2\pi/k$  that lies in the rotating plane with radius depending on  $n$ ,  $\alpha$  and  $\omega$ .*

An interesting situation appears when the integer closest to the angular velocity is not coprime with the number of bodies. In this case we prove that the minimal orbit is not circle anymore, as the following theorem states.

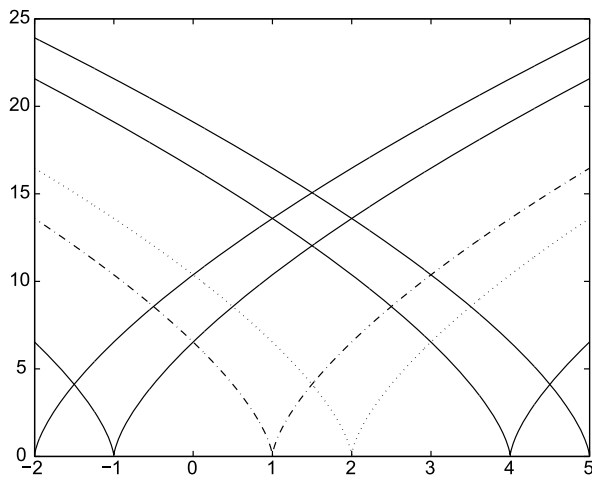
**Theorem 10** Take  $k \in \mathbb{N}$  and  $\text{g.c.d.}(k, n) = \tilde{k} > 1$ ,  $\tilde{k} \neq n$ . Then there exists  $\epsilon = \epsilon(\alpha, n, k) > 0$  such that if  $\omega \in (k - \epsilon, k + \epsilon) \setminus \{k\}$  the minimum of the action is attained on a planar  $2\pi$ -periodic orbit with winding number  $k$  which is not a relative equilibrium motion.

Also, it has been noticed that, for large number of bodies and angular velocities close to the half of an integer, the minimizer apparently is not anymore planar.

### Mountain Pass Solutions for the Choreographical 3-Body Problem

The discussion carried in the previous section shows that, as the angular velocity varies, the minimizer's shape must undergoes some transitions (for example it has to pass from relative equilibrium having different winding numbers). This scenario suggests the presence of other critical points, such as local minimizers or mountain pass. This was indeed discovered numerically in [12] and then proved by a computer assisted proof in [7].

To begin with, let us look at the following figure where the values of the action functional  $\mathcal{A}^\omega$  on the branches of circular orbits  $L_k^\omega$  are plotted:



The analysis of this picture suggests the presence of critical points different from the Lagrange motions. Indeed, let us take the angular velocity  $\omega = 1.5$ : in this case there are two distinct global minimizers, the uniform circular motions with minimal period  $2\pi$  and  $\pi$ , lying in the plane orthogonal to the rotation direction. This is a well known structure in Critical Point Theory, referred as the Mountain Pass geometry and gives the existence of a third critical point, provided the Palais–Smale condition is fulfilled, with an additional information on the Morse index. Next theorem fol-

lows from the application of the Mountain Pass Theorem to the action functional  $\mathcal{A}^{3/2}$ :

**Theorem 11** There exists a (possibly collision) critical point for the action functional  $\mathcal{A}^{3/2}$  with Morse index smaller than 1 and distinct from any Lagrange motion.

Once the existence of a Mountain Pass critical point was theoretically established, we studied its main properties in order to understand whether it belonged to some known families of periodic trajectories. To this aim we applied the bisection algorithm proposed in [13] to approximate the maximal of a locally optimal path joining the two strict global minimizers, finding in this a good numerical candidate. Of course, there could be a gap between the mountain pass solution whose existence is ensured by Theorem 11 and the numerical candidate found by applying the bisection algorithm. In order to fill this gap proved the existence of an actual solution very close to the numerical output of the Mountain Pass algorithm. The argument was based upon a fixed point principle and involved a rigorous computer assisted proof. As a consequence, we obtained the existence of a new branch of solution for the spatial 3-body problem (see Fig. 8).

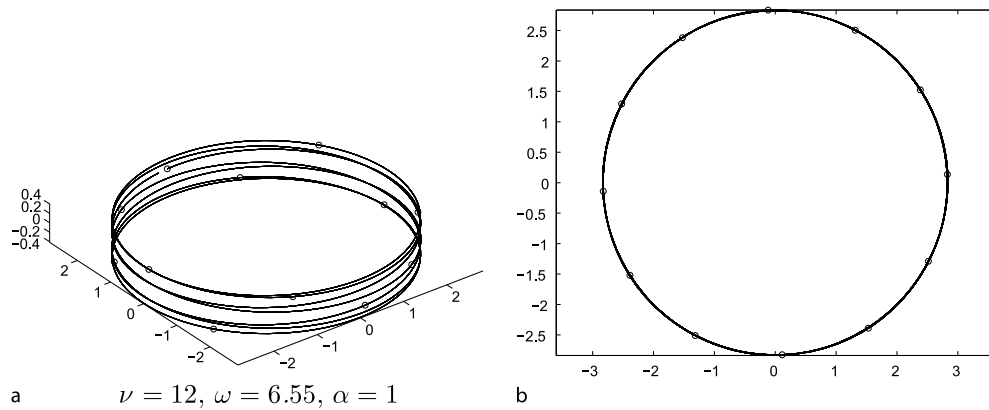
Here are some relevant features of the new solution: the orbit is not planar, its winding number with respect, for instance, to the line  $x = -0.2$ ,  $y = 0$  is 2 and it does not intersect itself. A natural question is whether this solution can be continued as a function of the parameter  $\omega$ . We were able to extend the numerical–rigorous argument to cover a full interval of values of the angular velocity, providing the existence of a full branch of solution.

**Theorem 12** There exists a smooth map  $B(\omega)$  giving the (locally unique, up to symmetries) branch of solution of the choreographical 3-body problem for all  $\omega \in [1, 2]$ , starting at the Mountain Pass solutions for  $\omega = 1.5$ .

A natural question is whether this mountain pass branch meets one of the known branches of choreographical periodic orbits: either one of the Lagrange or Marchal's  $P_{12}$  (described by Marchal in [62]) families. Surprisingly enough, it turns out from further numerical computation that this branch does not emanates from any of Lagrange motions  $L_1$  or  $L_2$ ; apparently, it emanates from the branch of  $P_{12}$  solutions. The details of the bifurcation diagram are depicted in Fig. 8.

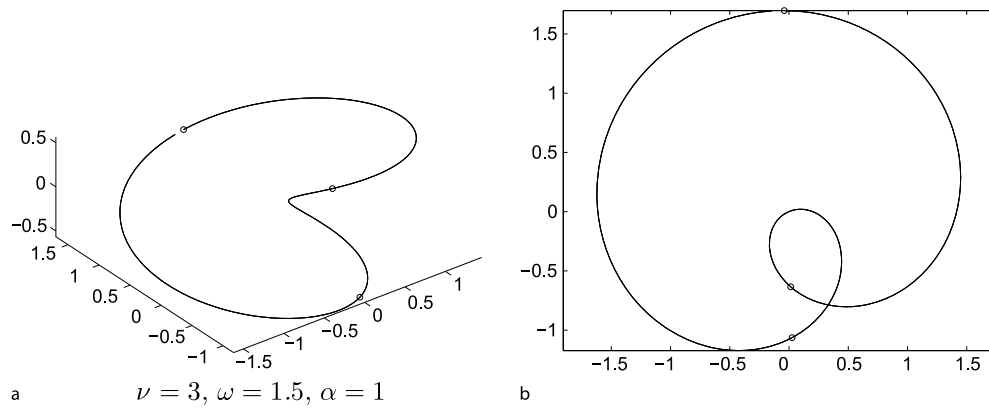
### Generalized Orbits and Singularities

The existence of a  $G$ -equivariant minimizer of the action is a simple consequence of the direct method of the Calculus of Variations. These trajectories, however, solve the associated differential equations only where they are far from



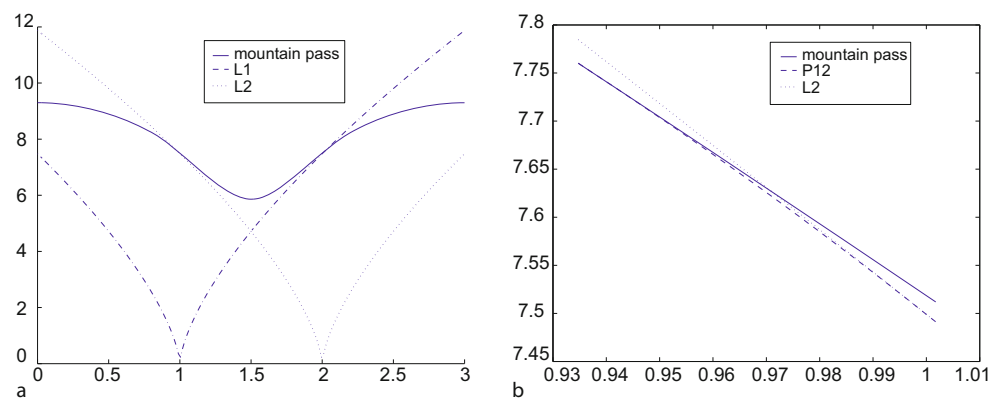
$n$ -Body Problem and Choreographies, Figure 19

Non planar minimizers of the action with angular velocities close to the half on an integer



$n$ -Body Problem and Choreographies, Figure 20

Mountain pass solution with angular velocities close to the half on an integer



$n$ -Body Problem and Choreographies, Figure 21

Action levels for the Lagrange and the mountain pass solution in the 3-body problem. On the  $x$ -axes the angular velocity varies in the interval  $[0, 3)$ . The *left picture* focuses on the bifurcation from the  $P_{12}$  family



the singular set. Indeed, generally speaking,  $G$ -equivariant minimizers may present singularities, even though the set of collision instants must clearly be of vanishing Lebesgue measure. A first natural question concerns the number of possible collision instants. Here below we follow [15].

**Definition 5** A path  $x: (a, b) \rightarrow \mathcal{X}$  is called a *locally minimizing solution* of the  $N$ -body problem if, for every  $t_0 \in (a, b)$ , there exists  $\delta > 0$  such that the restriction of  $x$  to  $[t_0 - \delta, t_0 + \delta]$  is a local minimizer for the action with the same boundary conditions. A path which is the uniform limit of a sequence of locally minimal solutions will be called a *generalized solution*.

We remark that:

- The definition can be modified in a obvious manner to include symmetries.
- Equivariant minimizers are generalized solutions.
- If the potential is of class  $C^2$  outside collisions then every non collision solution is a generalized solution.
- The generalized solutions possess an index: the minimal number of intervals  $I_j$  needed to cover  $(a, b)$  such that the restriction of  $x$  to  $I_j$  is a local minimizer for the action.
- There is a natural notion of maximal existence interval for generalized solutions, even without the unique extension property.

### Singularities and Collisions

Generally speaking we are dealing with systems of the form

$$M\ddot{x} = \nabla U(t, x), \quad t \in (a, b), \quad M_{ij} = m_i \delta_{ij},$$

where the potential  $U$  possesses a singular set  $(a, b) \times \Delta$ , in the sense that

$$[U0] \quad \lim_{x \rightarrow \Delta} U(t, x) = +\infty, \quad \text{uniformly in } t.$$

The set  $\Delta$  is clearly the *collision set*. We assume  $\Delta$  to be a *cone* in  $\mathbb{R}^{nd}$ :

$$x \in \Delta \implies \lambda x \in \Delta.$$

**Definition 6** We say that a generalized solution  $x$  on the interval  $(a, b)$ , has a *singularity* at  $t^* < +\infty$  if

$$\limsup_{t \rightarrow t^*} U(t, x(t)) = +\infty.$$

When  $t^* \in (a, b)$  we will say that  $x$  has an *interior singularity* at  $t = t^*$ , while when  $t^* = a$  or  $t^* = b$  (when finite) we will talk about a *boundary singularity*.

### The Theorems of Painlevé and Von Zeipel

In the usual mathematical language a classical solution  $x$  on the interval  $(a, b)$ , has a singularity at  $t^* < +\infty$  if it is not possible to extend  $x$  as a (classical) solution to a larger interval  $(a, t^* + \delta)$ . A classical results relates singularities of solutions with those of the potential.

**Theorem 13 (Painlevé's Theorem)** Let  $\tilde{x}$  be a classical solution for the  $n$ -body dynamical system on the interval  $[0, t^*)$ . If  $\tilde{x}$  has a singularity at  $t^* < +\infty$ , then

$$\lim_{t \rightarrow t^*} U(\tilde{x}(t)) = +\infty.$$

Painlevé's Theorem does not necessarily imply that a *collision* (i. e. that is a singularity such the configuration has a definite limit) occurs when there is a singularity at a finite time (on this subject we refer to [66,75,76]). The next result has been stated by Von Zeipel in 1908 and definitely proved by Sperling in 1970.

**Theorem 14 (Von Zeipel's Theorem)** If  $\tilde{x}$  is a classical solution for the  $n$ -body dynamical system on the interval  $(a, t^*)$  with a singularity at  $t^* < +\infty$  and

$$\lim_{t \rightarrow t^*} \|\tilde{x}(t)\| < +\infty,$$

then  $\tilde{x}(t)$  has a definite limit configuration  $x^*$  as  $t$  tends to  $t^*$ .

### Von Zeipel's Theorem and the Structure of the Collision Set

To the aim of extending the Von Zeipel's Theorem to our notion of generalized solutions, we need to introduce some assumptions on the potential  $U$  and its singular set  $\Delta$ .

$$\Delta = \bigcup_{\mu \in \mathcal{M}} V_\mu, \quad (5)$$

where the  $V_\mu$ 's are distinct linear subspaces of  $\mathbb{R}^k$  and  $\mathcal{M}$  is a finite set; observe that the set  $\Delta$  is a cone, as required before. We endow the family of the  $V_\mu$ 's with the inclusion partial ordering and we assume the family to be closed with respect to intersection (thus we are assuming that  $\mathcal{M}$  is a semilattice of linear subspaces of  $\mathbb{R}^k$ : it is the intersection semilattice generated by the arrangement of maximal subspaces  $V_\mu$ 's). With each  $\xi \in \Delta$  we associate

$$\mu(\xi) = \min\{\mu: \xi \in V_\mu\} \quad \text{i. e.}, \quad V_{\mu(\xi)} = \bigcap_{\xi \in V_\mu} V_\mu.$$

Fixed  $\mu \in \mathcal{M}$  we define the set of collision configurations satisfying

$$\Delta_\mu = \{\xi \in \Delta: \mu(\xi) = \mu\}$$

and we observe that this is an open subset of  $V_\mu$  and its closure  $\overline{\Delta_\mu}$  is  $V_\mu$ . We also notice that the map  $\xi \rightarrow \dim(V_{\mu(\xi)})$  is lower semicontinuous.

We denote by  $p_\mu$  the orthogonal projection onto  $V_\mu$  and we write

$$x = p_\mu(x) + w_\mu(x),$$

where, of course,  $w_\mu = \mathbb{I} - p_\mu$ .

We assume that, near the collision set, the potential depends, roughly, only on the projection orthogonal to the collision set: more precisely we assume

[U5] For every  $\xi \in \Delta$ , there is  $\varepsilon > 0$  such that  $U(t, x) - U(t, w_{\mu(\xi)}(x)) = W(t, x) \in C^1((a, b) \times B_\varepsilon(\xi))$ , where  $B_\varepsilon(\xi) = \{x: |x - \xi| < \varepsilon\}$ .

With these assumptions, we can extend then Von Zeipel's Theorem to generalized solutions:

**Theorem 15** *Let  $\bar{x}$  be a generalized solution on the bounded interval  $(a, b)$ . If  $\bar{x}$  is bounded on the whole interval  $(a, b)$  then the singularities of  $\bar{x}$  are collisions.*

## Asymptotic Estimates at Collisions

### One Side Conditions on the Potential and Its Radial Derivative

From now on, we require our potential  $U$  to satisfy some one-side homogeneity conditions:

[U1] there exists  $C_1 \geq 0$  such that for every  $(t, x) \in ((a, b) \times \mathbb{R}^{nd} \setminus \Delta)$

$$\left| \frac{\partial U}{\partial t}(t, x) \right| \leq C_1 U(t, x).$$

[U2] there exist  $\alpha \in (0, 2)$ ,  $\gamma > 0$  and  $C_2 \geq 0$  such that

$$\nabla U(t, x) \cdot x + \alpha U(t, x) \geq -C_2 |x|^\gamma U(t, x),$$

whenever  $|x|$  is small.

We observe that when  $U$  is homogeneous functions of degree  $-\alpha$ , the equality in condition (U2) is attained with  $C_2 = 0$ ; this assumption is satisfied also when we take into account potentials of the form  $U(t, x) = U_\alpha(x) + U_\beta(x)$  where  $U_\alpha$  is homogeneous of degree  $-\alpha$ ,  $U_\beta$  is homogeneous of degree  $-\beta$ ,  $0 < \beta < \alpha$ , and  $U_\alpha$  is positive.

### Isolatedness of Collisions Instants

In order to study the behavior of the solution at a collision, we first perform a translation in such a way that the

collision holds at the origin; next we introduce *polar coordinates*:

$$r = \sqrt{Mx \cdot x} = \sqrt{I}, \quad s = \frac{x}{r},$$

where  $s \in \mathcal{E} = \{x: I^2(x) = Mx \cdot x = 1\}$  belongs to the ellipsoid of all the configurations having unitary moment of inertia.

[U3<sub>h</sub>] there exists a function  $\tilde{U}$  defined on  $(a, b) \times (\mathcal{E} \setminus \Delta)$ , such that (on compact subsets of  $((a, b) \times \mathcal{E} \setminus \Delta)$ ):

$$\lim_{r \rightarrow 0} r^\alpha U(t, rs) = \tilde{U}(t, s);$$

[U3<sub>l</sub>] there exists a function  $\tilde{U}$  defined on  $(a, b) \times (\mathcal{E} \setminus \Delta)$ , such that (on compact subsets of  $((a, b) \times \mathcal{E} \setminus \Delta)$ ):

$$\lim_{|x| \rightarrow 0} [U(t, x) + M(t) \log |x|] = \tilde{U}(t, s);$$

The following regularity theorem is proved in [15] based on a suitable variant of Sundman's inequality and the asymptotic analysis of possible collisions outlined in the sequel.

**Theorem 16** *Let  $x: (a, b) \rightarrow \mathcal{X}$  be a generalized solution of the N-body problem, with a potential  $U$  satisfying [U0–U3]. Then collision instants are isolated in  $(a, b)$ . Furthermore, if  $(a, b)$  is the finite maximal extension interval of  $x$  and no escape in finite time occurs then the number of collision instants is finite.*

Some remarks are in order:

- A generalized solution does not solve the Euler–Lagrange equation in a distributional sense (the force field can not be locally integrable). Moreover its action needs not to be finite: one needs to prove it.
- a priori our solution can have a huge set of collision instants (more than countable, but null measure);
- there may be accumulation of partial collisions at a collision having more bodies;
- there is no a priori bound on the total action on the whole interval  $(a, b)$ , nor on the energy;
- we have very weak assumptions on the potential and only a one side inequality on the radial component
- the theorem extends to the logarithmic potentials.

### Conservation Laws

Even though they can not satisfy the differential equations in a distributional sense, generalized solutions satisfy a number of *conservation laws*.

**Theorem 17** *Let  $x$  be a generalized solutions on  $(a, b)$ . Then*

- The action  $\mathcal{A}(x, [a, b])$  on  $(a, b)$  is finite
- The energy  $h$  is bounded and belongs to the Sobolev space  $W^{1,1}((a, b), \mathbb{R})$ .
- Lagrange–Jacobi inequality holds in the sense of measures:

$$\frac{1}{2}\ddot{I}(\tilde{x}(t)) \geq 2h(t)(\tilde{x}(t)) + (2 - \alpha)U(t, \tilde{x}(t)) - C_2|\tilde{x}|^\gamma U(t, \tilde{x}), \quad \forall t \in (a, b).$$

- A monotonicity formula holds (extending Sundman's inequality): let us consider the angular energy

$$\Gamma_\alpha(r, s) := r^\alpha \left[ \frac{1}{2} r^2 |\dot{s}|^2 - U(t, rs) \right]$$

the  $\Gamma_\alpha$  is bounded variation and

$$\begin{aligned} \frac{d}{dt} \Gamma_\alpha(r, s) &\geq -\frac{2-\alpha}{2} r^{1+\alpha} \dot{r} |\dot{s}|^2 - C_1 r^\alpha U(t, rs) \\ &\quad + C_2 r^{\alpha+\gamma} \frac{\dot{r}}{r} U(t, rs). \end{aligned}$$

### Generalized Sundman–Sperling Estimates

The asymptotic analysis along a single collision (total or partial) trajectory goes back, in the classical case, to the works by Sundman ([84]), Wintner ([89]) and, in more recent years by Sperling, Pollard, Saari, Diacu and other authors (see for instance [40,47,75,76,79,83]). Now we are in a position to extend such estimate also to generalized solutions.

#### Theorem 18

- The following asymptotic estimates hold:

$$\begin{aligned} r &\sim (\kappa t)^{\frac{2}{2+\alpha}} \quad (r \sim |t| \sqrt{-\log(|t|)}) \\ K &\sim U \sim \frac{1}{4-2\alpha} \ddot{I} \sim \frac{2}{(2+\alpha)^2} \kappa^2 (\kappa t)^{\frac{-2\alpha}{2+\alpha}} \\ &\quad (\sim -\log |t|). \end{aligned}$$

- Let  $s$  be the normalized configuration of the colliding cluster  $s = x/r$ . Then

$$\begin{aligned} \lim_{t \rightarrow 0} r^2 |\dot{s}|^2 &= 0 \\ \lim_{t \rightarrow 0} U(s(t)) &= b < +\infty \end{aligned}$$

- Moreover there exists the angular blow-up, that is angular scaled family  $(s(\lambda t))_\lambda$  is precompact for the topology of uniform convergence on compact sets of  $\mathbb{R} \setminus \{0\}$ .

As a further consequence we have the vanishing of the total angular momentum.

In the same setting of the Theorem, assume that the potential  $U$  verifies the further assumption:

$$[U4] \lim_{r \rightarrow 0} r^{\alpha+1} \nabla_T U(t, x) = \nabla_T \tilde{U}(t, s).$$

Then we have

$$\lim_{t \rightarrow t_0} \text{dist}(C^b, s(t)) = \lim_{t \rightarrow t_0} \inf_{\tilde{s} \in C^b} |s(t) - \tilde{s}| = 0,$$

where  $C^b$  is the set of central configurations for  $\tilde{U}$  at level  $b$ .

### Dissipation and McGehee Coordinates

The main tool in proving the asymptotic estimates follow is the *monotonicity formula*, which is somehow equivalent to *Sundman's inequality*. To fix our mind, let us think to a homogeneous potential  $U_\alpha$ . A possible way to see this dissipation is to perform the change of variables (reminiscent of the McGehee coordinates):

$$\rho = r^{\frac{2-\alpha}{4}}, \quad \rho' = \frac{2-\alpha}{4} r^{-\frac{2+\alpha}{4}} r'$$

to obtain the action functional depending on  $(\rho, s)$ :

$$\begin{aligned} \mathcal{A}(\rho, s) &= \int_0^{\tau^*} \frac{1}{2} \left( \frac{4}{2-\alpha} \right)^2 (\rho')^2 \\ &\quad + \rho^2 \left( \frac{1}{2} |s'|^2 + U_\alpha(s) \right) - \lambda \rho^\beta d\tau, \end{aligned}$$

where  $\beta := 2(2+\alpha)/(2-\alpha) > 2$ . Here we have reparametrized the time as

$$dt = r^{\frac{2+\alpha}{2}} d\tau,$$

one proves that

$$\tau^* = \int_0^1 r^{-\frac{2+\alpha}{2}} dt = +\infty.$$

This is the coupling between a Duffing equation and the  $N$ -body angular system. When  $\rho$  is increasing, it acts as a viscosity coefficient on the angular Lagrangian.

A classical framework for the study of collisions is given by the McGehee coordinates [65] (here and below we assume, for simplicity of notations, all the masses be equal to one):

$$\begin{aligned} r &= |x| = I^{1/2} \\ s &= \frac{x}{r} \\ v &= r^{\alpha/2} (y \cdot s) \\ u &= r^{\alpha/2} (y - y \cdot s s). \end{aligned}$$

The equation of motions become (here  $'$  denotes differentiation with respect to the new time variable  $\tau$ ):

$$\begin{aligned} r' &= rv \\ v' &= \frac{\alpha}{2}v^2 + |u|^2 - \alpha U_\alpha(s) \\ s' &= u \\ u' &= \left(\frac{\alpha}{2} - 1\right)vu - |u|^2s + \alpha U_\alpha(s)s + \nabla U_\alpha(s). \end{aligned}$$

It is worthwhile noticing that the three last equations do not depend on  $r$  and hence are defined also for  $r = 0$ . In this context, the monotonicity formula means that  $v$  is a Lyapunov function for the system.

### Blow-ups

For every  $\lambda > 0$  let

$$x^\lambda(t) = \lambda^{-2/(2+\alpha)} x(\lambda t).$$

If  $\{\lambda_n\}_n$  is a sequence of positive real numbers such that  $s(\lambda_n)$  converges to a normalized configuration  $\bar{s}$ , then

$$\forall t \in (0, 1): \lim_{n \rightarrow \infty} s(\lambda_n t) = \lim_{n \rightarrow \infty} s(\lambda_n) = \bar{s}.$$

Hence the rescaled sequence will converge uniformly to the blow-up of  $x(t)$  relative to the colliding cluster  $\mathbf{k} \subset \mathbf{n}$  (in  $t = 0$ ).

- The blow-up  $\bar{x}$  is parabolic: where a *parabolic collision trajectory* for the cluster  $\mathbf{k}$  is the path

$$\bar{x}_i(t) = |t|^{2/(2+\alpha)} \xi_i, \quad i \in \mathbf{k}, \quad t \in \mathbb{R},$$

where  $\xi = (\xi_i)_{i \in \mathbf{k}}$  is a central configuration with  $k$  bodies.

**Proposition 8** *The sequences  $x^{\lambda_n}$  and  $(dx^{\lambda_n})/(dt)$  converge to the blow-up  $\bar{x}$  and its derivative  $\dot{\bar{x}}$  respectively, in the  $H^1$ -topology. Moreover  $\bar{x}$  is a minimizing trajectory in the sense of Morse.*

$$\int_0^T [\mathcal{L}(\bar{x} + \varphi) - \mathcal{L}(\bar{q}) \geq 0] dt.$$

for any compactly supported variation  $\varphi$ .

### Logarithmic Type Potentials

It is not possible to define a blow-up suitable for logarithmic type potentials. Indeed, the natural scaling should be

$\bar{x}^{\lambda_n}(t) := \lambda_n^{-1} \bar{x}(\lambda_n t)$ , which does not converge, since, by the asymptotics  $|x(t)| \simeq |t| \sqrt{-\log |t|}$ , we have:

$$\begin{aligned} \lim_{\lambda_n \rightarrow 0} |\bar{x}^{\lambda_n}(t)| &= \lim_{\lambda_n \rightarrow 0} \frac{r(\lambda_n t)}{\lambda_n t \sqrt{-2M(0) \log(\lambda_n t)}} \\ &\times t \sqrt{-2M(0) \log(\lambda_n t)} = +\infty \end{aligned}$$

for every  $t > 0$ .

On the other, hand, looking at the differential equation, the (right) blow-up should be:

$$\bar{q}(t) := t\bar{s}, \quad i \in \mathbf{k},$$

where  $\bar{s}$  is a central configuration for the system limit of a sequence  $s(\lambda_n)$  where  $(\lambda_n)_n$  is such that  $\lambda_n \rightarrow 0$ . This limiting function is the pointwise limit of the normalized sequence

$$\bar{x}^{\lambda_n}(t) := \frac{1}{\lambda_n \sqrt{-2M(0) \log \lambda_n}} \bar{x}(\lambda_n t).$$

Unfortunately this path is not locally minimal for the limiting problem, indeed since, the sequence  $(\bar{x}^{\lambda_n})_n$  converges to 0 as  $n$  tends to  $+\infty$ , and hence this blow-up minimizes only the kinetic part of the action functional.

### Absence of Collision for Locally Minimal Paths

As a matter of fact, solutions to the Newtonian  $n$ -body problem which are minimals for the action are, very likely, free of any collision. This fact was observed by the construction of suitable local variation arguments for the 2 and 3-body cases by Serra and Terracini [81,82]. The 4-body case was treated afterward by Dell'Antonio ([35], with a non completely rigorous argument) and then by A. Venturelli [87]. In general, the proof goes by the sake of the contradiction and involves the construction of a suitable variation that lowers the action in presence of a collision. A recent breakthrough in this direction is due to the neat idea, due to C. Marchal, [63], of averaging over a family of variations parametrized on a sphere. The method of averaged variations for Newtonian potentials has been outlined and exposed by Chenciner [25], and then fully proved and extended to  $\alpha$ -homogeneous potentials and various constrained minimization problems by Ferrario and Terracini in [52]. This technique can be used in many of the known cases to prove that minimizing trajectories are collisionless.

Of course, in some specific situations, other arguments can be useful, such as level estimates, on the infimum of the action on colliding paths [17,20,21,25]. However, these argument require global conditions on the potentials and

can not be applied in the present setting, where we work under local assumptions about the singularities.

We are in a position to prove the absence of collisions for locally minimal solutions when the potentials have quasi-homogeneous or logarithmic singularities. The first case is simpler, because one can take advantage the blow-up technique already exploited in [52]. On the other hand, when dealing with logarithmic potentials, the blow-up technique is no longer available and we conclude proving directly some averaging estimates that can be used to show the nonminimality of large classes of colliding motions. The following results are exposed in [15].

### Quasi-Homogeneous Potentials

Let  $\tilde{U}$  be the  $C^1$  function defined on  $(a, b) \times (\mathbb{R}^k \setminus \Delta)$  in the following way:

$$\tilde{U}(t, x) = |x|^{-\alpha} \lim_{r \rightarrow 0} r^\alpha \tilde{U}(t, rx/|x|).$$

With a slight abuse of notation, we denote  $\tilde{U}(x) = \tilde{U}(t^*, x)$ .  $\tilde{U}$  is homogeneous of degree  $-\alpha$ .

[U6] there is a 2-dimensional linear subspace of  $V_{\mu(\xi)}^\perp$ , say  $W$ , where  $\tilde{U}$  is rotationally invariant:

$$\tilde{U}(e^{i\theta} w) = \tilde{U}(w), \quad \forall w \in W, \quad \forall \theta \in [0, 2\pi];$$

[U7] for every  $x \in \mathbb{R}^k$  and  $\delta \in W$  there holds:

$$\begin{aligned} \tilde{U}(x + \delta) \leq \tilde{U} \left( \left( \frac{\tilde{U}(\pi_W(x))}{\tilde{U}(x)} \right)^{1/\alpha} \pi_W(x) \right. \\ \left. + \left( \frac{\tilde{U}(x)}{\tilde{U}(\pi_W(x))} \right)^{1/\alpha} \delta \right), \end{aligned}$$

where  $\pi_W$  denotes the orthogonal projection onto  $W$ .

**Theorem 19** In addition to [U0], [U1], [U2<sub>h</sub>], [U3<sub>h</sub>], [U4<sub>h</sub>], [U5], assume that, for all  $\xi \in \Delta$  [U6] and [U7] hold. Then generalized solutions do not have collisions at the time  $t^*$ .

**Remark 2** As our potential  $\tilde{U}$  is homogeneous of degree  $-\alpha$  the function

$$\varphi(x) = \tilde{U}^{-1/\alpha}(x)$$

is a non negative, homogeneous of degree one function, having now  $\Delta$  as zero set. In most of our applications  $\varphi$  will be indeed a quadratic form. Assume that  $\varphi^2$  splits in the following way:

$$\varphi^2(x) = K|\pi_W(x)|^2 + \varphi^2(\pi_{W^\perp}(x))$$

for some positive constant  $K$ . Then [U6] and [U7] are satisfied. Indeed, denoting  $w = \pi_W(x)$  and  $z = x - w$  we have, for every  $\delta \in W$ ,

$$\begin{aligned} \varphi^2(x + \delta) &= K|w + \delta|^2 + \varphi^2(z) \\ &= K \left| \frac{\varphi(x)}{\varphi(w)} w + \frac{\varphi(w)}{\varphi(x)} \delta \right|^2 + K \frac{\varphi^2(z)}{\varphi^2(x)} |\delta|^2 \\ &\geq K \left| \frac{\varphi(x)}{\varphi(w)} w + \frac{\varphi(w)}{\varphi(x)} \delta \right|^2 \\ &= \varphi^2 \left( \frac{\varphi(x)}{\varphi(w)} w + \frac{\varphi(w)}{\varphi(x)} \delta \right), \end{aligned}$$

which is obviously equivalent to [U7].

**Proposition 9** Assume  $\tilde{U}(x) = Q^{-\alpha/2}(x)$  for some non negative quadratic form  $Q(x) = \langle Ax, x \rangle$ . Then assumptions [U6] and [U7] are satisfied whenever  $W$  is included in an eigenspace of  $A$  associated with a multiple eigenvalue.

Given two potentials satisfying [U6] and [U7] for a common subspace  $W$ , their sum enjoys the same properties. On the other hand, the class of potentials satisfying [U6] and [U7] is not stable with respect to the sum of potentials. In order to deal with a class of potentials which is closed with respect to the sum, we introduce the following variant of the last Theorem.

**Theorem 20** In addition to [U0], [U1], [U2<sub>h</sub>], [U3<sub>h</sub>], [U4<sub>h</sub>], [U5], assume that  $\tilde{U}$  has the form

$$\tilde{U}(x) = \sum_{v=1}^N \frac{K_v}{(\text{dist}(x, V_v))^\alpha},$$

where  $K_v$  are positive constants and  $V_v$  is a family of linear subspaces, with  $\text{codim}(V_v) \geq 2$ , for every  $v = 1, \dots, N$ . Then locally minimizing trajectories do not have collisions at the time  $t^*$ .

These two theorems extend to logarithmic potentials.

### Neumann Boundary Conditions and G-equivariant Minimizers

Our analysis allows to prove that minimizers to the fixed-ends (Bolza) problems are free of collisions: indeed all the variations of our class have compact support. However, other type of boundary conditions (generalized Neumann) can be treated in the same way. Indeed, consider a trajectory which is a (local) minimizer of the action among all paths satisfying the boundary conditions

$$x(0) \in X^0, \quad x(T) \in X^1,$$



where  $X^0$  and  $X^1$  are two given linear subspaces of the configuration space. Consider a (locally) minimizing path  $\bar{x}$ : of course it has not interior collisions. In order to exclude boundary collisions we have ensure that the class of variations preserve the boundary conditions. This can be achieved by imposing assumptions [U6] and [U7] to be fulfilled also by the restriction of the potential to the boundary subspaces  $X^i$ . The analysis of boundary conditions was a key point in the paper [52], where symmetric periodic trajectories were constructed by reflections about given subspaces. Our results can be used to prove the absence of collisions also for  $G$ -equivariant (local) minimizers, provided the group  $G$  satisfies the Rotating Circle Property defined in Definition 3. Hence, existence of  $G$ -equivariant collisionless periodic solutions can be proved for the wide class of symmetry groups described in [14,52], for a much larger class of interacting potentials, including quasi-homogeneous and logarithmic ones. On the other hand, our results can be applied to prove that  $G$ -equivariant minimals are collisionless for many relevant symmetry groups violating the rotating circle property, such as the groups of rotations recently introduced in [50,51].

### The Standard Variation

In order to prove that local minimizers are free of collisions we are going to make use of the following class of variations:

**Definition 7** The standard variation associated to  $\delta \in \Xi$  and  $T$  is defined as

$$v^\delta(t) = \begin{cases} \delta & \text{if } 0 \leq |t| \leq T - |\delta| \\ (T - t) \frac{\delta}{|\delta|} & \text{if } T - |\delta| \leq |t| \leq T \\ 0 & \text{if } |t| \geq T \end{cases}.$$

Our next goal is to find a standard variation  $v^\delta$  such that the trajectory  $q + v^\delta$  does not have a collision at  $t = 0$  and

$$\Delta \mathcal{A} := \int_{-\infty}^{+\infty} [\mathcal{L}_k(q + v^\delta) - \mathcal{L}_k(q)] dt < 0.$$

Let us introduce the function

$$S(\xi, \delta) = \int_0^{+\infty} \left( \frac{1}{|\xi t^{2/(2+\alpha)} - \delta|^\alpha} - \frac{1}{|\xi t^{2/(2+\alpha)}|^\alpha} \right) dt,$$

where  $\xi, \delta \in \mathbb{R}^2$ . We first estimate the variation of the action as  $\delta \rightarrow 0$ :

**Theorem 21** Let  $q = \{q\}_i = \{t^{2/(2+\alpha)} \xi_i\}$ ,  $i = 1, \dots, k$  be a parabolic collision trajectory and  $v^\delta$  a standard variation.

Then, as  $\delta \rightarrow 0$

$$\Delta \mathcal{A} = 2|\delta|^{1-\alpha/2} \sum_{\substack{i < j \\ i, j \in k}} m_i m_j S\left(\xi_i - \xi_j, \frac{\delta_i - \delta_j}{|\delta|}\right) + O(|\delta|).$$

We observe that

$$S(\lambda \xi, \mu \delta) = |\lambda|^{-1-\alpha/2} |\mu|^{1-\alpha/2} S(\xi, \delta)$$

and hence the sign of  $S$  depends on the angle between  $\xi$  and  $\delta$ . Let

$$\Phi(\vartheta) = \int_0^{+\infty} \frac{1}{\left(t^{\frac{4}{\alpha+2}} - 2 \cos \vartheta t^{\frac{2}{\alpha+2}} + 1\right)^{\alpha/2}} - \frac{1}{t^{\frac{2\alpha}{\alpha+2}}} dt, \quad \alpha \in (0, 2).$$

$\Phi(\theta)$  represents the potential differential needed for displacing the colliding particle from zero to  $e^{i\theta}$ . Expanding, we can express  $\Phi(\theta)$  as a hypergeometric function:

$$\begin{aligned} \Phi(\vartheta) &= \frac{\alpha(\alpha+2)}{2} \left\{ \frac{1}{\alpha-2} \beta \left( \frac{\alpha+2}{4}, \frac{\alpha+2}{4} \right) \right. \\ &\quad + \frac{1}{\alpha} \sum_{k=1}^{+\infty} \binom{-\alpha/2}{k} (-1)^k 2^{k-1} (\cos \vartheta)^k \beta \\ &\quad \times \left( \frac{\alpha}{4} - \frac{1}{2} + \frac{k}{2}, \frac{\alpha}{4} + \frac{1}{2} + \frac{k}{2} \right) \left. \right\}. \end{aligned}$$

### Some Properties of $\Phi$

The value of  $\Phi(\theta)$  ranges from  $+\infty$  to some negative value, depending on  $\alpha$ . However, thanks to some harmonic analysis one can prove that suitable averages are always negative: the first inequality is particularly useful for dealing with reflected triple collisions from the Lagrange central configuration:

$$\Phi\left(\frac{2\pi}{3} + \gamma\right) + \Phi\left(\frac{2\pi}{3} - \gamma\right) < 0, \quad \forall \gamma \in [0, \pi/2].$$

A key remark was made by Christian Marchal: being the Newton potential a harmonic map averaging it on a sphere results in a truncation in the interior. In fact, is not so much a matter of harmonicity. A crucial estimate was proved in [FT] about the averages of  $\Phi$  on circles:

**Theorem 22** For every  $\alpha > 0$ ,  $\xi \in \mathbb{R}^3 \setminus \{0\}$  and for every circle  $\mathbb{S} \subset \mathbb{R}^d$  with center in 0,

$$\begin{aligned} \tilde{S}(\xi, \mathbb{S}) &= \frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} S(\xi, \delta) d\delta \\ &= |\xi|^{-1-\alpha/2} |\delta|^{1-\alpha/2} \frac{1}{2\pi} \int_0^{2\pi} \Phi(\theta) d\theta < 0. \end{aligned}$$

Consider  $\xi = x_i - x_j$  and  $\delta$  ranging in a circle. Then the above inequality implies the following principle, a generalization of Marcha's statement in [25]:

"it is more convenient (from the point of view of the integral of the potential on the time line) to replace one of the point particles with a homogeneous circle of same mass and fixed radius which is moving keeping its center in the position of the original particle."

### Future Directions

This article mainly focuses on the variational approach to the search of selected trajectories to the  $n$ -body problem. In our examples, the masses can very well be equal: hence the problem can not be regarded as a small perturbation of a simpler one and a full picture of the dynamics is out of reach. In contrast, the planetary  $n$ -body problem deals with systems where one of the bodies (a "star") is much heavier than the others (the "planets") and can be seen as a perturbation of a decoupled systems of one center problems. A small parameter  $\varepsilon$  represents the order of the ratio between the mass of the planets and that of the star. The system is then called *nearly integrable* as it can be associated with a Hamiltonian of the form:

$$H(I, \varphi) = h(I) + \varepsilon f(I, \varphi),$$

where  $I \in \mathbb{R}^N$  and  $\varphi \in \mathbb{T}^N$  ( $N$  is the number of degree of freedom) and  $\varepsilon$  is a small parameter. In the three body problem we have  $N = 4$ , but the integrable limit possesses motions lying on  $\mathbb{T}^2$  (the product of two Keplerian orbits) so *the integrable limit depends on less action variables than the number of degrees of freedom*. For this reason the system is *properly degenerate*. The integrable limit possesses only quasi-periodic motions: in [56], the question whether these motions survive for positive values of  $\varepsilon$  is settled. KAM Theory deals with the problem of persistence of such invariant tori but can not be directly applied to the planetary  $n$ -body problem, for its degeneracy. Nevertheless, the existence of invariant tori has been recently achieved in the planar and spatial three body problem [8,16,18,57,78] and in the planetary many body problem [19,48].

The future researches will extend these results to some non perturbative settings, finding both regular motions, such as periodic, quasi-periodic or quasi-periodic trajectories and irregular, chaotic ones, through the application of suitable variational methods taking into account of collision trajectories.

## Bibliography

### Primary Literature

1. Albouy A (1996) The symmetric central configurations of four equal masses. *Contemp Math* 198:131–135
2. Albouy A (1998) Chenciner A, Le problème des  $n$  corps et les distances mutuelles. *Invent Math* 131:151–184
3. Ambrosetti A, Coti Zelati V (1993) Periodic solutions of singular Lagrangian systems. In: *Progress in Nonlinear Differential Equations and their Applications*, vol 10. Birkhäuser Boston Inc., Boston
4. Ambrosetti A, Coti Zelati V (1994) Non-collision periodic solutions for a class of symmetric 3-body type problems. *Topol Meth Nonlin Anal* 3(2):197–207
5. Ambrosetti A, Rabinowitz PH (1973) Dual variational methods in critical point theory and applications. *J Funct Anal* 14:349–381
6. Arioli G, Gazzola F, Terracini S (2000) Minimization properties of Hill's orbits and applications to some  $N$ -body problems. *Ann Inst H Poincaré Anal Non Linéaire* 17(5):617–650
7. Arioli G, Barutello V, Terracini S (2006) A new branch of mountain pass solutions for the choreographical 3-body problem. *Commun Math Phys* 268(5):439–463
8. Arnold VI (1963) Small denominators and problems of stability of motions in classical and celestial mechanics. *Uspehi Naut Nauk* 18(6):91–192
9. Bahri A, Rabinowitz PH (1991) Periodic solutions of Hamiltonian systems of 3-body type. *Ann Inst H Poincaré Anal Non Linéaire* 8(6):561–649
10. Barutello V (2004) On the  $n$ -body problem, Ph. D thesis, Università di Milano-Bicocca, available at <http://www.matapp.unimib.it/dottorato/>
11. Barutello V, Secchi S (2008) Morse index properties of colliding solutions to the  $n$ -body problem. *Arxiv:math/0609837*, *Annales de l'Institut Henri Poincaré (C) Non Linear Anal* 25:539–565
12. Barutello V, Terracini S (2004) Action minimizing orbits in the  $n$ -body problem with choreography constraint. *Nonlinearity* 17:2015–2039
13. Barutello V, Terracini S (2007) A bisection algorithm for the numerical Mountain Pass, *NoDEA* 14:527–539
14. Barutello V, Ferrario DL, Terracini S (2008) Symmetry groups of the planar 3-body problem and action-minimizing trajectories. *Arch Rat Mech Anal* 190:189–226
15. Barutello V, Ferrario DL, Terracini S (2008) On the singularities of generalized solutions to the  $N$ -body problem. *Int Math Res Notices* 2008:rnn069–78
16. Berti M, Biasco L, Valdinoci E (2004) Periodic orbits close to invariant tori and applications to the three-body problem. *Ann Scuola Norm Sup Pisa Cl Sci 5 vol III*:87–138
17. Bessi U, Coti Zelati V (1991) Symmetries and noncollision closed orbits for planar  $N$ -body-type problems. *Nonlin Anal* 16(6):587–598
18. Biasco L, Chierchia L, Valdinoci E (2003) Elliptic two dimensional invariant tori for the planetary three-body problem. *Arch Rat Mech Anal* 170:91–135
19. Biasco L, Chierchia L, Valdinoci E (2006)  $N$ -dimensional elliptic invariant tori for the planetary  $(N + 1)$ -body problem. *SIAMJ Math Anal* 37:1560–1588
20. Chen K-C (2001) Action-minimizing orbits in the parallelogram

- four-body problem with equal masses. *Arch Rat Mech Anal* 158:293–318
21. Chen K-C (2001) On Chenciner–Montgomery's orbit in the three-body problem. *Discrete Contin Dyn Syst* 7(1):85–90
22. Chen K-C (2003) Binary decompositions for planar *n*-body problems and symmetric periodic solutions. *Arch Rat Mech Anal* 170(3):247–276
23. Chen K-C (2003) Variational methods on periodic and quasi-periodic solutions for the *N*-body problem. *Ergodic Theory Dyn Syst* 23(6):1691–1715
24. Chenciner A (2002) Action minimizing periodic orbits in the Newtonian *n*-body problem *Celestial Mechanics*, dedicated to Don Saari. *Contemp Math* 292:71–90
25. Chenciner A (2002) Action minimizing solutions of the newtonian *n*-body problem: from homology to symmetry. *ICM, Peking*
26. Chenciner A (2002) Simple non-planar periodic solutions of the *n*-body problem. In: *Proceedings of the NDDS Conference, Kyoto*
27. Chenciner A, Desolneux N (1998) Minima de l'intégrale d'action et équilibre relatifs de *n* corps. *C R Acad Sci Paris Sér I* 326:1209–1212; Correction in: (1998) *C R Acad Sci Paris Sér I* 327:193
28. Chenciner A, Féjóz J (2005) L'équation aux variations verticales d'un équilibre relatif comme source de nouvelles solutions périodiques du problème des *N* corps. *C R Math. Acad. Sci. Paris* 340(8):593–598
29. Chenciner A, Montgomery R (2000) A remarkable periodic solution of the three body problem in the case of equal masses. *Ann Math* 152(3):881–901
30. Chenciner A, Venturelli A (2000) Minima de l'intégrale d'action du problème Newtonien de 4 corps de masses égales dans  $\mathbb{R}^3$ : orbites "hip-hop". *Celest Mech* 77:139–152
31. Chenciner A, Féjóz J, Montgomery R (2005) Rotating eights. I. The three  $\Gamma$  families. *Nonlinearity* 18(3):1407–1424
32. Chenciner A, Gerver J, Montgomery R, Simó C (2001) Simple choreographies of *N* bodies: a preliminary study. In: *Geometry, Mechanics and Dynamics*. Springer, New York, pp 287–308
33. Degiovanni M, Giannoni F (1988) Dynamical systems with newtonian type potentials. *Ann. Scuola Norm Sup Pisa, Ser IV* 15:467–494
34. Degiovanni M, Giannoni F, Marino A (1987) Dynamical systems with newtonian type potentials. *Atti Accad Naz Lincei Rend Cl Sci Fis Mat Natur Ser* 8(81):271–278
35. Dell'Antonio G (1998) Non-collision periodic solutions of the *N*-body system. *NoDEA Nonlinear Differ Equ Appl* 5:1 117–136
36. Devaney RL (1978) Collision orbits in the anisotropic Kepler problem. *Invent Math* 45(3):221–251
37. Devaney RL (1978) Nonregularizability of the anisotropic Kepler problem. *J Differ Equ* 29(2):252–268
38. Devaney RL (1980) Triple collision in the planar isosceles three-body problem. *Invent Math* 60(3):249–267
39. Devaney RL (1981) Singularities in classical mechanical systems. in: *Ergodic theory and dynamical systems, I*. College Park, Md., 1979–80, vol 10 of *Progr Math*. Birkhäuser, Boston, pp 211–333
40. Diacu F (1992) Regularization of partial collisions in the *N*-body problem. *Differ Integral Equ* 5(1):103–136
41. Diacu F (1993) Painlevé's conjecture. *Math Intell* 15(2):6–12
42. Diacu F (1996) Near-collision dynamics for particle systems with quasihomogeneous potentials. *J Differ Equ* 128(1):58–77
43. Diacu F (2002) Singularities of the *N*-body problem. in: *Classical and celestial mechanics*. Princeton Univ. Press, Princeton, pp 35–62, Recife, 1993/1999
44. Diacu F, Santoprete M (2004) On the global dynamics of the anisotropic Manev problem. *Phys D* 194(1–2):75–94
45. Diacu F, Pérez-Chavela E, Santoprete M (2006) Central configurations and total collisions for quasihomogeneous *n*-body problems. *Nonlinear Anal* 65(7):1425–1439
46. Diacu F, Pérez-Chavela E, Santoprete M (2005) The Kepler problem with anisotropic perturbations. *J Math Phys* 46(7):072701, 21
47. ElBialy MS (1990) Collision singularities in celestial mechanics. *SIAM J Math Anal* 21(6):1563–1593
48. Féjóz J (2004) Démonstration du "théorème d'Arnold" sur la stabilité du système planétaire (d'après Michel Herman). (french) [Proof of "Arnold's theorem" on the stability of a planetary system (following Michel Herman)], *Ergodic Theory Dyn Syst* 24(5):1521–1582
49. Ferrario DL (2002) Symmetric periodic orbits for the *n*-body problem: some preliminary results, Preprint of the Max-Planck-Institut für Mathematik MPI-2002-79
50. Ferrario DL (2007) Transitive decomposition of symmetry groups for the *n*-body problem. *Adv Math* 213:763–784
51. Ferrario DL (2006) Symmetry groups and non-planar collisionless action-minimizing solutions of the three-body problem in three-dimensional space. *Arch Rat Mech Anal* 179(3):389–412
52. Ferrario D, Terracini S (2004) On the existence of collisionless equivariant minimizers for the classical *n*-body problem. *Invent Math* 155(2):305–362
53. The GAP Group (2002) GAP – Groups, Algorithms, and Programming, Version 4.3, <http://www.gap-system.org>
54. Gordon WB (1975) Conservative dynamical systems involving strong forces. *Trans Am Math Soc* 204:113–135
55. Gordon WB (1977) A minimizing property of Keplerian orbits. *Am J Math* 99(5):961–971
56. Jefferys WH, Moser J (1966) Quasi-periodic solutions for the three-body problem. *Celest Mech Dyn Astron* J 71:508–578
57. Laskar J, Robutel P (1995) Stability of the planetary three-body problem, I: Expansion of the planetary Hamiltonian. *Celest Mech Dyn Astron* 62(3):193–217
58. Levi Civita T (1918) Sur la régularization du problème des trois corps. *Acta Math* 42:99–144
59. Majer P, Terracini S (1993) Periodic solutions to some problems of *n*-body type. *Arch Rat Mech Anal* 124(4):381–404
60. Majer P, Terracini S (1995) On the existence of infinitely many periodic solutions to some problems of *n*-body type. *Commun Pure Appl Math* 48(4):449–470
61. Majer P, Terracini S (1995) Multiple periodic solutions to some *n*-body type problems via a collision index Variational methods in nonlinear analysis (Erice, 1992). Gordon and Breach, Basel, pp 245–262
62. Marchal C (2000) The family  $P_{12}$  of the three-body problem – the simplest family of periodic orbits, with twelve symmetries per period. *Celest Mech Dyn Astron* 78(1–4):279–298; (2001) New developments in the dynamics of planetary systems. *Badhofgastein*, 2000
63. Marchal C (2002) How the method of minimization of action avoids singularities. *Celest Mech Dyn Astron* 83:325–353
64. Mather JN, McGehee R (1974) Solutions of the collinear four body problem which become unbounded in finite time. In: *Dynamical systems, theory and applications*. Rencontres, Battelle

- Res Inst, Seattle, Wash., pp 573–597. Lecture Notes in Phys., vol 38. Springer, Berlin, (1975)
65. McGehee R (1974) Triple collision in the collinear three-body problem. *Invent Math* 27:191–227
  66. McGehee R (1986) von Zeipel's theorem on singularities in celestial mechanics. *Exposition Math* 4(4):335–345
  67. Moeckel R (1990) On central configurations, *Math Zeit* 205:499–517
  68. Moeckel R (1987) Some qualitative features of the three-body problem. in: *Hamiltonian dynamical systems*. (Boulder, 1987) vol 81 of *Contemp Math*, pp 1–22. Am Math Soc, Providence RI, 1988
  69. Montgomery R (1998) The  $N$ -body problem, the braid group, and action-minimizing periodic solutions. *Nonlinearity* 11(2):363–376
  70. Montgomery R (1999) Action spectrum and collisions in the planar three-body problem. In: *Celestial Mechanics*. (Evanston, 1999) vol 292 of *Contemp Math Am Math Soc*, Providence RI, 2002, pp 173–184
  71. Moore C (1993) Braids in classical dynamics. *Phys Rev Lett* 70(24):3675–3679
  72. Pacella F (1987) Central configurations and the equivariant Morse theory. *Arch Rat Mech* 97:59–74
  73. Palais RS (1979) The principle of symmetric criticality. *Commun Math Phys* 69:19–30
  74. Poincaré H (1896) Sur les solutions périodiques et le principe de moindre action. *C R Acad Sci Paris, Sér I Math* 123:915–918
  75. Pollard H, Saari DG (1968) Singularities of the  $n$ -body problem I. *Arch Rat Mech Anal* 30:263–269
  76. Pollard H, Saari DG (1970) Singularities of the  $n$ -body problem II. In: *Inequalities II*. Academic Press, New York, pp 255–259 (Proc. Second Sympos, US Air Force Acad, Colo, 1967)
  77. Riahi H (1999) Study of the critical points at infinity arising from the failure of the Palais-Smale condition for  $n$ -body type problems. *Mem Am Math Soc* 138:658, viii+112
  78. 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
  79. Saari DG (1972/73) Singularities and collisions of Newtonian gravitational systems. *Arch Rat Mech Anal* 49:311–320
  80. Sbano L (1998) The topology of the planar three-body problem with zero total angular momentum and the existence of periodic orbits. *Nonlinearity* 11(3):641–658
  81. Serra E, Terracini S (1992) Collisionless periodic solutions to some three-body problems. *Arch Rat Mech Anal* 120(4):305–325
  82. Serra E, Terracini S (1994) Noncollision solutions to some singular minimization problems with Keplerian-like potentials. *Nonlinear Anal* 22(1):45–62
  83. Sperling HJ (1970) On the real singularities of the  $N$ -body problem. *J Reine Angew Math* 245:15–40
  84. Sundman KF (1913) Mémoire sur le problème des trois corps. *Acta Math* 36:105–179
  85. Terracini S, Venturelli A (2007) Symmetric trajectories for the  $2N$ -body problem with equal masses. *Arch Rat Mech Anal* 184(3):465–493
  86. Venturelli A (2001) Une caractérisation variationnelle des solutions de Lagrange du problème plan des trois corps, *C R Acad Sci Paris, Sér I Math* 332(7):641–644
  87. Venturelli A (2002) Application de la minimisation de l'action au Problème des  $N$  corps dans le plan et dans l'espace. Thesis. University Paris VII
  88. Wang Q (1991) The global solution of the  $n$ -body problem, *Celest Mech Dyn Astron* 50(1):7388
  89. Wintner A (1941) The analytical foundations of celestial mechanics. Princeton Mathematical Series, vol 5. Princeton University Press, Princeton
  90. Xia Z (1992) The existence of non collision singularities in newtonian systems. *Ann Math* 135:411–468
  91. Von Zeipel H (1908) Sur les singularités du problème des  $n$  corps. *Ark Math Astr Fys* 4:1–4

## Books and Reviews

- Arnold VI, Kozlov V, Neishtadt A (2006) Mathematical aspects of classical and celestial mechanics [Dynamical systems. III]. 3rd edn. In: *Encyclopaedia of Mathematical Sciences*, vol 3. Springer, Berlin, xiv and p 518, translated from the Russian original by Khukhro E
- Diacu F, Holmes P (1996) Celestial encounters. The origins of chaos and stability. Princeton University Press, Princeton
- Meyer, Kenneth R (1999) Periodic solutions of the  $N$ -body problem, *Lecture Notes in Mathematics*, 1719. Springer, Berlin
- Moser J (1973) Stable and random motions in dynamical systems, With special emphasis on celestial mechanics, Hermann Weyl Lectures, the Institute for Advanced Study, Princeton NJ, *Annals of Mathematics Studies*, No. 77. Princeton University Press, Princeton
- Pollard H (1976) Celestial mechanics, Carus Mathematical Monographs, 18. Mathematical Association of America, Washington
- Saari D (2005) Collisions, rings, and other Newtonian  $N$ -body problems, CBMS Regional Conference Series in Mathematics, 104. American Mathematical Society, Providence RI, Washington, Published for the Conference Board of the Mathematical Sciences
- Siegel CL, Moser JK (1995) Lectures on celestial mechanics. Classics in Mathematics. Springer, Berlin, Translated from the German by Kalme CI, Reprint of the 1971 translation
- Stiefel EL, Scheifele G (1971) Linear and regular celestial mechanics. Perturbed two-body motion, numerical methods, canonical theory. *Die Grundlehren der mathematischen Wissenschaften*, Band 174. Springer, New York, Heidelberg

## Nekhoroshev Theory

LAURENT NIEDERMAN<sup>1,2</sup>

<sup>1</sup> Topologie et Dynamique – UMR 8628 du CNRS, Université Paris, Paris, France

<sup>2</sup> Astronomie et Systèmes Dynamiques – UMR 8028 du CNRS, IMCCE, Paris, France

## Article Outline

Glossary

Definition of the Subject

Introduction



Exponential Stability of Constant Frequency Systems  
 Nekhoroshev Theory (Global Stability)  
 Applications  
 Future Directions  
 Appendix: An Example of Divergence  
 Without Small Denominators  
 Bibliography

## Glossary

**Quasi integrable Hamiltonian** A Hamiltonian is quasi integrable if it is close to another Hamiltonian whose associated system is integrable by quadrature. Here, we will consider real analytic Hamiltonians on a domain  $\mathcal{D}$  which admit a holomorphic extension on a complex strip  $\mathcal{D}_{\mathbb{C}}$  around  $\mathcal{D}$  and the closeness to the integrable Hamiltonian is measured with the supremum norm  $\|\cdot\|_{\infty}$  over  $\mathcal{D}_{\mathbb{C}}$ .

**Exponentially stable Hamiltonian** An integrable Hamiltonian governs a system which admits a collection of first integral. We say that an integrable Hamiltonian  $h$  is exponentially stable if for any small enough Hamiltonian perturbation of  $h$ , the solutions of the perturbed system are at least defined over timescales which are exponentially long with respect to the inverse of the size of the perturbation. Moreover, the first integrals of the integrable system should remain nearly constant along the solutions of the perturbed system over the same amount of time.

**Nekhoroshev theorem (1977)** There exists a *generic set* of real analytic integrable Hamiltonians which are *exponential stable*.

## Definition of the Subject

We only know explicitly the solutions of the Hamiltonian systems which are integrable by quadrature. Unfortunately these integrable Hamiltonians are exceptional but many physical problems can be studied by a Hamiltonian system which differs from an integrable one by a small perturbation. One of the most famous is the motions of the planets around the Sun which can be seen as a perturbation of the integrable system associated to the motion of noninteracting points around a fixed attracting center. Poincaré [70] considered the study of these quasi integrable Hamiltonian systems as the *Problème général de la dynamique*. This question was tackled with the Hamiltonian perturbation theories which were introduced at the end of the nineteenth century precisely to study the planetary motions (see ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#)).

Significant theorems concerning the mathematical justification of these methods were not put forward until the 1950s and later. A cornerstone of these results is the Kolmogorov–Arnold–Moser (KAM) theory which states that under suitable nondegeneracy and smoothness assumptions, for a small enough perturbation most of the solutions of a nearly integrable Hamiltonian system are quasi periodic, hence they are stable over infinite times. More accurately, KAM theory provides a set of large measure of invariant tori which support stable and global solutions of the perturbed system. But this set of stable quasi periodic solutions has a complicated topology: it is a Cantor set nowhere dense and without interior points. Numerically, it is extremely difficult to determine if a given solution is quasi periodic or not.

Moreover, for a  $n$  degrees of freedom Hamiltonian system (hence a  $2n$ -dimensional system) KAM theory provides  $n$ -dimensional tori. If  $n = 2$ , these two-dimensional invariant tori divide the three-dimensional energy level, therefore the solutions of the perturbed system are global and bounded over *infinite* times but an arbitrary large drift of the orbits is still possible for  $n \geq 3$ . Actually, Arnold [1] proposed examples of quasi integrable Hamiltonian systems where an arbitrary large instability occurs for an arbitrary small perturbation. This is known as “Arnold diffusion”.

Thus, results of stability for quasi integrable Hamiltonian systems which are valid for an open set of initial condition can only be proved over *finite* times.

The first theorems of effective stability over finite but very long times were proved by Moser [58] and Littlewood [45] around an elliptic equilibrium point in an Hamiltonian system. Extending this kind of result in a general framework, Nekhoroshev [62,63] proved in 1977 a fundamental theorem of global stability which completes KAM theory and states that, for a *generic set* of integrable Hamiltonian, Arnold diffusion can only occur over exponentially long times with respect to the inverse of the size of the perturbation.

These theories have been applied extensively in celestial mechanics but also to study molecular dynamics, beam dynamics, billiards, geometric numerical integrators, stability of certain solutions of nonlinear PDE (Schrodinger, wave, ...).

## Introduction

We first recall that the canonical equations with a given scalar function  $\mathcal{H}$  (the Hamiltonian) which is differentiable on an open set  $\mathcal{D} \subset \mathbb{R}^{2n}$  equipped with the symplectic form of Liouville  $\omega = \sum_i dp_i \wedge dq_i$  for  $(q_1, \dots, q_n) \in$



$\mathbb{R}^n$  and their conjugate coordinates  $(p_1, \dots, p_n) \in \mathbb{R}^n$  can be written:

$$\dot{p} = -\partial_q \mathcal{H}(p, q); \quad \dot{q} = \partial_p \mathcal{H}(p, q).$$

Actually, this kind of system can be defined over any symplectic manifold.

Moreover, a diffeomorphism  $\Phi$  is *symplectic* or *canonical* if it preserves Liouville form:  $\Phi^* \omega = \omega$ . Especially, for a Hamiltonian  $X$ , the flow  $\Phi_X^t$  linked to the canonical system governed by  $X$  is a symplectic transformation over its domain of definition [54]. Such a diffeomorphism preserves the Hamilton equations: in the new variables  $(p, q) = \Phi(P, Q)$  the transformed system is still canonical with the initial Hamiltonian  $K(P, Q) = \mathcal{H} \circ \Phi(P, Q)$ .

### Integrable and Quasi-integrable Hamiltonian Systems

According to the theorem of Liouville–Arnold, under general topological conditions, a Hamiltonian system integrable by quadrature can be reduced to a system defined over  $\Omega \times \mathbb{T}^n$  for an open set  $\Omega \subset \mathbb{R}^n$  and the  $n$ -dimensional torus  $\mathbb{T}^n$  with a Hamiltonian which does not depend on the  $n$  angles. The new variables  $(I, \theta) \in \Omega \times \mathbb{T}^n$  are called *actions-angle* variables and the system is linked to a Hamiltonian  $K(I)$ , hence the equations take the trivial form  $\dot{I} = 0, \dot{\theta} = \nabla K(I)$  where  $\nabla K$  is the gradient of  $K$  and we obtain quasi-periodic motions of frequencies  $\nabla K(I)$ .

If a small Hamiltonian perturbation is added to an integrable system, then in action-angle variables the considered system is governed by

$$\mathcal{H}(\varepsilon, I, \theta) = h(I) + \varepsilon f(I, \theta) \quad \text{where } \mathcal{H} \in C^\omega(\Omega \times \mathbb{T}^n, \mathbb{R}) \quad (1)$$

and the equations become:

$$\dot{I} = -\varepsilon \partial_\theta f(I, \theta); \quad \dot{\theta} = \nabla h(I) + \varepsilon \partial_I f(I, \theta). \quad (2)$$

Hence, the variables are separated into two groups: the fast variables and the other variables which evolve slowly (the actions but also the angles with zero frequencies), over times of order  $1/\varepsilon$  their evolution may be of order 1.

### Averaging Principle

The *averaging principle* consists of the replacement of the initial system by its time average along the unperturbed flow  $\Phi_h^t$  linked to  $h$  which means that we consider the averaged Hamiltonian:

$$\langle \mathcal{H} \rangle(I, \theta) = h(I) + \varepsilon \langle f \rangle(I, \theta) \\ \text{with } \langle f \rangle(I, \theta) = \lim_{t \rightarrow \infty} \left( \frac{1}{t} \int_0^t f(I, \theta + s \nabla h(I)) ds \right).$$

Actually, this average depends on the commensurability relations which are satisfied by the components of vector  $\nabla h(I)$ .

More accurately, to a submodule  $\mathcal{M} \subset \mathbb{Z}^n$ , we associate its *resonant zone*:

$$\mathcal{Z}_{\mathcal{M}} = \{I \in \mathbb{R}^n \text{ such that } k \cdot \nabla h(I) = 0 \\ \text{if and only if } k \in \mathcal{M}\}. \quad (3)$$

Let  $r \in \{0, \dots, n-1\}$  be the rank of  $\mathcal{M}$ , there exists a unimodular matrix  $R \in \text{SL}(n, \mathbb{Z})$  such that  $\omega \in \mathcal{M}^\perp$  if and only if the first  $r$  lines of  $R\omega$  are zeros. Hence, the symplectic transformation  $\phi = R\theta$  and  $I = {}^t R^{-1} J$  defined over  $\Omega \times \mathbb{T}^n$  yields the new Hamiltonian  $\tilde{h}(J) + \varepsilon \tilde{f}(J, \phi)$  where  $J = (J_1, J_2) \in \mathbb{R}^r \times \mathbb{R}^{n-r}$  and  $\phi = (\phi_1, \phi_2) \in \mathbb{T}^r \times \mathbb{T}^{n-r}$  such that  $\nabla \tilde{h}(J) = (0, \omega(J))$  when  ${}^t R^{-1} J \in \mathcal{Z}_{\mathcal{M}}$ . Moreover, the second component  $\omega(J) \in \mathbb{R}^{n-r}$  does not satisfy any commensurability relation.

Equivalently, on the set  $\mathcal{Z}_{\mathcal{M}}$ , we can consider the torus  $\mathbb{T}^n$  as the product  $\mathbb{T}^r \times \mathbb{T}^{n-r}$  where the unperturbed flow is *constant* over  $\mathbb{T}^r$  and *ergodic* over  $\mathbb{T}^{n-r}$ .

Hence, at infinity, the time average  $\langle \tilde{f} \rangle(J, \phi)$  is equal to the space average:

$$\langle \tilde{f} \rangle(J_1, J_2, \phi_1) = \left( \frac{1}{2\pi} \right)^{n-r} \iint_{\mathbb{T}^{n-r}} \tilde{f}(J_1, J_2, \phi_1, \phi_2) d\phi_2, \quad (4)$$

and the averaged system involves only variables which evolve slowly. For instance, to determine the solutions numerically, we only need to take a step of integration of order  $1/\varepsilon$  [41].

Actually, in the expression (4), we have only made a *partial averaging* by removing  $n-r$  angles. Hence, the considered system is simpler but, for a generic perturbation, it is *only* for the zero module  $\mathcal{M} = \{0\}$  that we obtain an integrable Hamiltonian.

It should be noticed that, for an arbitrary module  $\mathcal{M}$ , the solutions of the average system can be unbounded. For instance, the Hamiltonian  $\mathcal{H}(I_1, I_2, \theta_1, \theta_2) = I_1 I_2 + \varepsilon \sin(\theta_2)$  is averaged with respect to the module  $\mathcal{M} = \mathbb{Z} \mathbf{i}$  with  $\mathbf{i} = (1, 0)$  since  $\theta_1$  does not appear in the perturbation and it admits the unbounded solution  $(I_1(t), I_2(t), \theta_1(t), \theta_2(t)) = (0, -\varepsilon t, \varepsilon t^2/2, 0)$  which starts from the origin  $(0, 0, 0, 0)$  at  $t = 0$ . One can also notice that we have a maximal speed of drift of the action  $I_2$  with respect to the size of the perturbation.

On the other hand, a key observation for the proof of Nekhoroshev theorem is the fact that for an arbitrary module  $\mathcal{M}$ , the canonical equations ensure that the variations of the actions under the averaged vector field with respect

to  $\mathcal{M}$  are located in the *subspace spanned by  $\mathcal{M}$* . This point will be specified in Sect. “[The Initial Statement](#)”, it will be the only place in Nekhoroshev’s proof where the canonical form of the equations is used.

### Hamiltonian Perturbation Theory

The averaging principle is based on the idea that the oscillating terms discarded in averaging cause only small oscillations which are superimposed to the solutions of the averaged system. In order to prove the validity of the averaging principle, one should check that any solutions of the perturbed system remain close to the solution of the averaged system with the same initial condition. Especially, this will be the case if one finds a canonical transformation  $\varepsilon$ -close to identity which transforms the perturbed Hamiltonian to its average. Hence we are reduced to a problem of normal form where one looks for a convenient system of coordinates which gives the simplest possible form to the considered system.

Here, we consider the transformation  $\Phi_X^1$  given by the time 1 flow of the Hamiltonian system governed by  $X(I, \theta) = \varepsilon X_1(I, \theta)$ .

With Taylor formula, the transformed Hamiltonian  $\mathcal{H} \circ \Phi_X^1$  admits the following expansion with respect to  $\varepsilon$ :

$$\mathcal{H} \circ \Phi_X^1 = h + \varepsilon(f + \nabla h(I) \cdot \partial_\theta X_1(I, \theta)) + \mathcal{O}(\varepsilon^2),$$

and in order to obtain  $\mathcal{H} \circ \Phi_X^1 = h(I) + \varepsilon \langle f \rangle$  one must solve:

$$\nabla h(I) \cdot \partial_\theta X_1(I, \theta) = -f + \langle f \rangle, \quad (5)$$

which is the *homological equation*, this is the central equation of perturbation theory.

Since we are in the analytic setting, the function  $f$  admits the expansion

$$\sum_{k \in \mathbb{Z}^n} f_k(I) \exp(ik\theta).$$

Hence, for an averaging with respect to a resonant module  $\mathcal{M}$  around the resonant zone  $\mathcal{Z}_{\mathcal{M}}$  linked to  $\mathcal{M}$ , the homological equation admits the formal solution:

$$X_1(I, \theta) = \sum_{k \notin \mathcal{M}} \frac{f_k(I)}{i(k \cdot \nabla h(I))} \exp(ik\theta), \quad (6)$$

thus one obtains a transformation which normalizes the Hamiltonian at first order in  $\varepsilon$ .

In the same way, one can *formally* eliminate the fast angles at all orders by looking at a transformation generated by a Hamiltonian  $X(I, \theta) = \sum_{n \geq 1} \varepsilon^n X_n(I, \theta)$ . Indeed, the same type of homological equation appears at all order to determine  $X_n$  for  $n > 1$ .

This is the *Linstedt method*.

With the previous construction, it is obvious that the normalizing transformation admits an expansion which is divergent if the denominators  $k \cdot \nabla h(I)$  for  $k \notin \mathcal{M}$  are too close to zero on the considered domain in the action space. This is the well-known problem of the *small denominators* which was emphasized by Poincaré [70] in his celebrated theorem about nonexistence of analytic first integrals for a generic quasi integrable Hamiltonian.

It is less known that even without small denominators, the classical perturbation theory can yield divergent expansions. This is the problem of the *great multipliers* according to Poincaré terminology which come from the successive differentiations.

This problem is presented in the next subsection.

### Exponential Stability of Constant Frequency Systems

#### The Case of a Single Frequency System

The normalization of an analytic quasi-integrable Hamiltonian system with only *one fast* phase is one of the main problems in perturbation theory. This question appears naturally to compute the time of approximate conservation of the adiabatic invariants [17,49]. It has been accurately studied in [60] and [73] and we will focus our attention on this case where the phenomenon of divergence without small denominators appears in its simplest setting.

Indeed, as in the Sect. “[Hamiltonian Perturbation Theory](#)”, one can build *formally* the Hamiltonian  $X(I, \theta) = \sum_{n \geq 1} \varepsilon^n X_n(I, \theta)$  which generates a normalizing symplectic transformation and eliminates the fast angle in the perturbation. But, for a generic analytic quasi-integrable Hamiltonian, it can be shown [60,73] that perturbation theory yields a *Gevrey-2* normalizing transformation over  $U \times \mathbb{T}^n$  (i. e.:  $\mathcal{T} \in C^\infty(U \times \mathbb{T}^n)$  with  $\|\partial^k \mathcal{T}\|_\infty \leq CM^{2|k|} k!^2$  where  $C, M$  are positive constants and  $k = (k_1, \dots, k_{2n}) \in \mathbb{N}^{2n}$ ;  $|k| = |k_1| + \dots + |k_{2n}|$ ;  $k! = k_1! \dots k_{2n}!$ ) such that the initial perturbed Hamiltonian  $\mathcal{H} = h + \varepsilon f$  is transformed into an integrable Hamiltonian  $h_\varepsilon(\tilde{I})$  with a one parameter family  $h_\varepsilon$  of scalar functions Gevrey-2 over  $U$ .

Hence, the normalizing transformations are usually divergent. On the other hand, by general properties of Gevrey functions (see Sect. 3.3 in [53], and the references therein) if one considers the transformation generated by the truncated expansion

$$\sum_{n=1}^N \varepsilon^n X_n(I, \theta)$$

obtained after  $N$  steps of perturbation theory, then the transformed Hamiltonian is normalized up to a remainder of size  $\varepsilon^{N+1}N!$ .

In the appendix, we will study an example of a quasi-integrable Hamiltonian where this latter estimate cannot be improved and where the source of divergence of the normalizing transformation which comes from the successive differentiations in the construction can be emphasized.

We see that the remainder of size  $\varepsilon^{N+1}N!$  decreases rapidly before increasing to infinity, following Poincaré [70] this is a “convergent expansion according to astronomers” and a “divergent expansion according to geometers”.

Now, we can use the process of “summation at the smallest term”: for a fixed  $\varepsilon > 0$ , one obtains an optimal normalization with a truncation at order  $N$  such that

$$\|\partial_{\theta_2} X_{N-1}\|_{\infty} \simeq \varepsilon \|\partial_{\theta_2} X_N\|_{\infty} \quad \text{which yields } N = E(1/\varepsilon).$$

Finally, the Stirling formula yields the size of the remainder:  $\sqrt{2\pi\varepsilon} \exp(-1/\varepsilon)$  which is exponentially small with respect to the inverse of the size of the perturbation.

More generally, Marco and Sauzin [53] have proved in the same setting that starting from a Gevrey- $\alpha$  Hamiltonian ( $\|\partial^k \mathcal{H}(I, \theta)\|_{\infty} \leq CM^{\alpha|k|}(k!)^{\alpha}$ ), one can build a normalizing transformation which is Gevrey- $\alpha+1$  and these estimates cannot be improved usually but the previous construction is still possible. Indeed, one can still make a summation at the smallest term and obtain a canonical change of coordinates which normalizes the Hamiltonian up to an exponentially small remainder with respect to the size of the perturbation.

In the case where the averaged Hamiltonian is *integrable*, according to the mean value theorem, the speed of drift of the normalized action variables is at most *exponentially slow*. Since the size of the normalizing transformation is of order  $\varepsilon$ , the initial actions admit at most a drift of size  $\varepsilon$  over an *exponentially long time*.

### The Case of a Strongly Nonresonant Constant Frequency System

Systems with constant frequencies, hence  $h(I) = \omega \cdot I$  for some constant vector  $\omega \in \mathbb{R}^n$ , appear when we consider small nonlinear interactions of linear oscillatory systems or the action of quasi-periodic perturbations on linear oscillatory systems. In any case, the considered Hamiltonian can be written:  $\mathcal{H} = \omega \cdot I + f(I, \theta)$  with a small function  $f \in C^{\omega}(\Omega \times \mathbb{T}^n, \mathbb{R})$  where  $\Omega$  is an open set in  $\mathbb{R}^n$ .

Moreover, we assume here that the frequency  $\omega$  is a  $(\gamma, \tau)$ -Diophantine vector for some positive constants  $\gamma$  and  $\tau$ , hence  $\omega \in \Omega_{\gamma, \tau}$ :

$$\Omega_{\gamma, \tau} = \left\{ \omega \in \mathbb{R}^n \text{ such that } |k \cdot \omega| \geq \frac{\gamma}{\|k\|_{\infty}^{\tau}} \right. \\ \left. \text{for all } k \in \mathbb{Z}^n \setminus \{(0, \dots, 0)\} \right\}. \quad (7)$$

We recall that the measure of the complementary set of  $\Omega_{\gamma, \tau}$  is of order  $\mathcal{O}(\gamma)$  for  $\tau > n - 1$ .

Under these assumptions, one can prove [13,31,45,71] that for a small enough analytic perturbation, the action variables of the unperturbed problem become quasi-integrals of the perturbed system over exponentially long times, more specifically:

**Theorem 1 ([13,31,45,71])** Consider a Hamiltonian  $\omega \cdot I + f(I, \theta)$  real analytic over a domain  $\mathcal{U} \times \mathbb{T}^n \subset \mathbb{R}^n \times \mathbb{T}^n$  which admit a holomorphic extension on a complex strip of width  $\rho > 0$  around  $\mathcal{U} \times \mathbb{T}^n$  in  $\mathbb{C}^{2n}$ .

The supremum norm for a holomorphic function on this complex strip is denoted  $\|\cdot\|_{\rho}$ .

There exists positive constants  $C_1, C_2, C_3, C_4$  which depend only on  $\gamma, \tau, \rho, n$  such that if  $\varepsilon = \|f\|_{\rho} < C_1\gamma$ , an arbitrary solution  $(I(t), \varphi(t))$  of the perturbed system associated to  $\omega \cdot I + f(I, \theta)$  with an initial action  $I(t_0) \in \mathcal{U}$  is defined at least over an exponentially long time and satisfies:

$$\|I(t) - I(0)\| \leq C_2\varepsilon \quad \text{if } |t| \leq C_3 \exp(C_4\varepsilon^{-\frac{1}{1+\tau}}). \quad (8)$$

The proof is based on the existence of a normalizing transformation up to an exponentially small error. This is possible since we have lower bounds on the small denominators and the growth of the coefficients in the normalizing expansion is reduced to a combinatorial problem.

Finally, since the averaged Hamiltonian is integrable, the speed of drift of the action variables is at most exponentially slow.

### Nekhoroshev Theory (Global Stability)

#### The Initial Statement

Thirty years ago, Nekhoroshev [62,63] stated a *global* result of stability which is valid for a *generic* set of integrable Hamiltonian. Especially, we don't have anymore a control on the small denominators as in the previous section but we have to handle the resonant zones. Nekhoroshev's reasonings allow one to prove a global result of stability *inde-*

pendent of the arithmetical properties of the unperturbed frequencies by taking into account the *geometry of the integrable system*. This is really a change of perspective with respect to the previous results. The key ingredient is to find a suitable property of the integrable Hamiltonian, namely the property of *steepness* introduced in the sequel, which ensures that a drift of the actions in the averaged system with respect to a module  $\mathcal{M} \subset \mathbb{Z}^n$  leads to an *escape* of the resonant zone  $\mathcal{Z}_{\mathcal{M}}$ .

More specifically, Nekhoroshev proved global results of stability over open sets of the following type:

**Definition 2 (exponential stability)** Consider an open set  $\Omega \subset \mathbb{R}^n$ , an analytic integrable Hamiltonian  $h: \Omega \rightarrow \mathbb{R}$  and action-angle variables  $(I, \varphi) \in \Omega \times \mathbb{T}^n$  where  $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ .

For an arbitrary  $\rho > 0$ , let  $\mathcal{O}_\rho$  be the space of analytic functions over a complex neighborhood  $\Omega_\rho \subset \mathbb{C}^{2n}$  of size  $\rho$  around  $\Omega \times \mathbb{T}^n$  equipped with the supremum norm  $\|\cdot\|_\rho$  over  $\Omega_\rho$ .

We say that the Hamiltonian  $h$  is exponentially stable over an open set  $\tilde{\Omega} \subset \Omega$  if there exists positive constants  $\rho, C_1, C_2, a, b$  and  $\varepsilon_0$  which depend only on  $h$  and  $\tilde{\Omega}$  such that:

- i)  $h \in \mathcal{O}_\rho$ .
- ii) For any function  $\mathcal{H}(I, \varphi) \in \mathcal{O}_\rho$  such that  $\|\mathcal{H} - h\|_\rho = \varepsilon < \varepsilon_0$ , an arbitrary solution  $(I(t), \varphi(t))$  of the Hamiltonian system associated to  $\mathcal{H}$  with an initial action  $I(t_0)$  in  $\tilde{\Omega}$  is defined over a time  $\exp(C_2/\varepsilon^a)$  and satisfies:

$$\|I(t) - I(t_0)\| \leq C_1 \varepsilon^b \quad \text{for } |t - t_0| \leq \exp(C_2/\varepsilon^a), \quad (9)$$

$a$  and  $b$  are called stability exponents.

**Remark** Along the same lines, the previous definition can be extended to an integrable Hamiltonian in the Gevrey class (see [53]).

Hence, for a small enough perturbation, the action variables of the unperturbed problem become *quasi integrals* of the perturbed system over exponentially long times.

In order to introduce the problem, we begin by a typical example of *non-exponentially* stable integrable Hamiltonian:  $h(I_1, I_2) = I_1 I_2$ . Indeed, the perturbed system governed by  $h(I_1, I_2) + \varepsilon \sin(\theta_2)$  admits the unbounded solution  $(I_1(t), I_2(t), \theta_1(t), \theta_2(t)) = (0, \varepsilon t, \varepsilon t^2/2, 0)$  which starts from the origin  $(0, 0, 0, 0)$  at  $t = 0$ , hence a drift of the actions  $(I_1(t), I_2(t))$  on a segment of length 1 occurs over a timespan of order  $1/\varepsilon$ .

The important feature in this example which has to be avoided in order to ensure exponential stability is the fact

that the gradient  $\nabla h(I_1, 0)$  remains orthogonal to the first axis. Equivalently, the gradient of the restriction of  $h$  on this first axis is identically zero.

Nekhoroshev [61,62,63] introduced the class of *steep* functions where this problem is avoided. The property of steepness is a quantitative condition of transversality for a real valued function differentiable over an open set  $\Omega \subset \mathbb{R}^n$  which involves *all* the affine subspaces which intersect  $\Omega$ . Actually, steepness can be characterized by the following simple geometric criterion proved thanks to theorems of real subanalytic geometry [67]:

**Theorem 3 ([67])** A real analytic scalar function without critical points is steep if and only if its restriction to any proper affine subspace admits only isolated critical points.

This is an extension of a previous similar result in the holomorphic case [42].

In this setting, Nekhoroshev proved the following:

**Theorem 4 ([62,63])** If the integrable Hamiltonian  $h$  is real analytic, does not admit critical points, is nondegenerate ( $|\nabla^2 h(I)| \neq 0$  for any  $I \in \Omega$ ) and steep then  $h$  is exponentially stable.

The set of steep functions is *generic* among sufficiently smooth functions. For instance, we have seen that the function  $xy$  is not steep but it can easily be shown that  $xy + x^3$  is steep. Actually, a given function can be transformed into a steep function by adding higher order terms [61,62]. It can be noticed that the (quasi-)convex functions are the *steepest* functions since their restrictions to any affine subspaces admit at most *one* critical point which is also nondegenerate.

The original proof of Nekhoroshev is *global*. It is based on a covering of the action space in open sets with controlled resonance properties where one can build resonant normal forms (i. e.: where only resonant harmonics are retained) up to an exponentially small remainder. The averaged Hamiltonian is not necessarily integrable but, thanks to the steepness of the integrable Hamiltonian, if a drift of the normalized actions occurs then it can only lead to a zone associated to resonances of *lower multiplicity* than the initial one (i. e.: the resonant module admits a lower dimension). Eventually, after a short distance the orbits reach a resonance-free area (i. e.: the Fourier expansion of the normalized perturbation admits only nonresonant harmonics up to an exponentially small remainder). Then, the local normal form is integrable and yields the confinement of the action variables over the desired amount of time.



### Improved Versions of Nekhoroshev Theorem

The articles of Nekhoroshev remained largely unnoticed in the western countries until Benettin, Galgani and Giorgilli [12] rewrote and clarified the initial proof in the convex case.

Benettin and Gallavotti [13] proved that under an assumption of (quasi) convexity of the unperturbed Hamiltonian, the proofs of these theorems can be simplified. Indeed, after an averaging as in the steep case, the quasi convexity and the *energy conservation* ensure that the normalized Hamiltonian is an *approximate Liapunov function* over exponentially long time intervals. This allows one to confine the actions in the initial set where the considered orbit was located (we do not have to consider a drift over resonant areas of different multiplicity as in the original proof). Hence, the construction of a *single* normal form is enough to confine the actions in the convex case.

Following this idea, Lochak [46,47] has significantly simplified the proof of Nekhoroshev estimates for the convex quasi integrable Hamiltonians. His reasonings are based on normalization around the periodic orbits of the integrable Hamiltonian which represent the worst type of resonances. Using convexity, Lochak obtains open sets around the periodic orbits which hold exponential stability. Then, Dirichlet theorem about *simultaneous Diophantine approximation* ensures that these open sets recover the whole action space and yield the global result.

A remarkable feature of this proof is the fact that *improved estimates* can be obtained in the vicinity of resonances thanks to the relative abundance of periodic orbits in these areas. More specifically, periodicity corresponds to  $n - 1$  commensurability relations and we have already several commensurability relations at the resonances hence Dirichlet theorem can be applied on a lower dimensional space with better rates of approximation [46,47]. These improvements are important to extend Nekhoroshev estimates for large systems or infinite dimensional systems, they also fit with the speed of drift of the action variables in examples of unstable quasi integrable Hamiltonian (these points will be discussed in the sequel).

It can also be noticed that averaging along the periodic orbits of the integrable system is exactly a *one phase* averaging without small denominators.

Toward sharp estimates, Lochak–Neishtadt [50] and Pöschel [71] have independently obtained the following:

**Theorem 5 ([50,71])** *If the integrable Hamiltonian  $h$  is real analytic, does not admit critical points and convex over a domain  $\Omega \subset \mathbb{R}^n$  then  $h$  is exponentially stable over  $\Omega$  with the global exponents  $a = b = 1/2n$ .*

*Moreover, around the resonant zones linked to a module of rank  $m < n$ , the integrable Hamiltonian  $h$  is exponentially stable with the improved exponents  $a = b = 1/2(n - m)$ .*

The proof in [50], explicitly derived in [51], relies on Lochak periodic orbits method together with a refined procedure of averaging due to Neishtadt [60]. In [71], the original scheme of Nekhoroshev is combined with a refined study of the geometry of resonances which gives an accurate partition of the action space in open sets where the action variables are confined and also Neishtadt's averaging procedure is used. Pöschel's study of the geometry of resonances should also be important in the study of Arnold diffusion.

This value of the time exponent ( $a = 1/2n$ ) is expected to be *optimal* in the convex case according to heuristic reasonings of Chirikov [22], see also [46] on the speed of drift of Arnold diffusion. Actually, Marco–Sauzin [53] in the Gevrey cases and Marco–Lochak [52] in the analytic case have essentially proved the optimality of the improved exponent  $a = 1/2(n - 2)$  in the doubly resonant zones starting from an example of unstable quasi integrable Hamiltonian given by Michel Herman.

On the other hand, the previous studies except the original one of Nekhoroshev do not cover the cases of a time-dependent perturbation or a perturbed steep integrable Hamiltonian, despite their importance in physics. For instance, a time periodic perturbation of a convex Hamiltonian can be reduced to the time-independent perturbation of a quasi convex Hamiltonian in the extended phase space, but this is not the case for a general time-dependent perturbation, where energy conservation cannot be used. This problem has been studied in the light of Nekhoroshev theory by Giorgilli and Zehnder [34] in connection with the dynamics of a particle in a time-dependent potential with a high kinetic energy.

A new general study of the stability of steep integrable Hamiltonians has been carried out in [66]. This proof of stability relies on the mechanism of Nekhoroshev since we analyze the dynamics around resonances of any multiplicity and use local resonant normal forms but the original *global* construction is substituted with a *local* construction along each trajectory of the perturbed system. This construction is based on the approximation of the frequencies  $\nabla h(I(t))$  at certain times by rational vectors thanks to Dirichlet theorem of simultaneous Diophantine approximation as in Lochak's proof of exponential stability in the convex case. This allows significant simplifications with respect to Nekhoroshev's original proof. Moreover, the results of Lochak and Pöschel are generalized for the steep



case since the exponents of stability derived in [66] give back the exponents  $a = b = 1/2n$  in the particular case of convex integrable Hamiltonians.

In [68], Nekhoroshev estimates are proved for perturbations of integrable Hamiltonians which satisfy a *strictly weaker* condition of nondegeneracy than the initial condition of steepness. Indeed, the lack of steepness allows a drift of the action variables around the resonant zones. But, due to the exponential decay of the Fourier coefficients in the expansion of the Hamiltonian vector field, such a drift would be extremely slow if one consider resonant zones linked to a module spanned by integer vectors of large length. Thanks to this property, Morbidelli and Guzzo [57] have observed that the Hamiltonian  $h(I_1, I_2) = I_1^2 - \delta I_2^2$  where  $\delta$  is the square of a Diophantine number is *non steep* but nevertheless  $h$  is *exponentially stable* since its isotropic directions are the lines spanned by  $(1, \pm\sqrt{\delta})$  which are “far” from the lines with rational slopes. This phenomenon has been studied numerically for the quadratic integrable Hamiltonians in [39].

Starting from this observation, a general weak condition of steepness which involves only the affine subspaces spanned by *integer* vectors has been stated in [68] with a complete proof of exponential stability in this setting.

The point in this refinement lies in the fact that it allows one to exhibit a *generic* class of real analytic integrable Hamiltonians which are exponentially stable with *fixed* exponents of stability  $a$  and  $b$  while Nekhoroshev original theory provides a generic set of exponentially stable integrable Hamiltonians but with exponents of stability which can be *arbitrarily small* [68].

More specifically, we consider genericity in a measure theoretical sense since:

**Theorem 6 ([68])** *Consider an arbitrary real analytic integrable Hamiltonian  $h$  defined on a neighborhood of the closed ball  $\overline{B}_R^{(n)}$  of radius  $R$  centered at the origin in  $\mathbb{R}^n$ .*

*For almost any  $\Omega \in \mathbb{R}^n$ , the integrable Hamiltonian  $h_\Omega(x) = h(I) - \Omega \cdot I$  is exponentially stable with the exponents:*

$$a = \frac{b}{2 + n^2} \quad \text{and} \quad b = \frac{1}{2(1 + 2^n n)}.$$

Finally, all these results can be generalized for quasi integrable *symplectic mappings* thanks to a theorem of inclusion of an analytic symplectic diffeomorphism into the flow linked to a real analytic Hamiltonian up to an exponentially small accuracy ([43,75] or [37] for a direct proof).

## KAM Stability, Exponential Stability, Nekhoroshev Stability

A last point which should be emphasized is the link between KAM stability, exponential stability and Nekhoroshev stability which are cornerstones in the study of stability of analytic quasi integrable Hamiltonian systems.

A first problem is the stability of the solutions in a neighborhood of a Lagrangian invariant torus over which an analytic Hamiltonian vectorfield induces a linear flow of frequency  $\omega$ . Then the considered Hamiltonian can be written as  $\mathcal{H}(I, \theta) = \omega \cdot I + \mathcal{F}(I, \theta)$  in action-angle variables where the perturbation  $\mathcal{F}(I, \theta)$  is analytic in a neighborhood of the origin and starts at order *two* in actions.

In the case of a KAM torus, the frequency is strongly nonresonant (Diophantine). With a suitable rescaling, the expansion of the Hamiltonian takes the form considered in Theorem 1 and the exponential estimates of stability (9) are valid.

In the general case (where the frequency can satisfy resonances of low order), the previous procedure cannot be applied. On the other hand, if one assumes that there are no resonances of order lower or equal to *four*:

$$\forall k \in \mathbb{Z}^n \setminus \{0\} \quad \text{such that} \quad |k| = |k_1| + \dots + |k_n| \leq 4 \quad \text{then} \quad |k \cdot \omega| \neq 0,$$

it is possible to perform a Birkhoff's normalization which reduces the studied Hamiltonian to:

$$\widetilde{\mathcal{H}}(\widetilde{I}, \widetilde{\theta}) = \omega \cdot \widetilde{I} + \widetilde{Q}(\widetilde{I}) + \widetilde{\mathcal{F}}(\widetilde{I}, \widetilde{\theta}), \quad (10)$$

where  $\widetilde{Q}$  is a quadratic form (the torsion) and  $\widetilde{\mathcal{F}}(\widetilde{I}, \widetilde{\theta}) = \mathcal{O}_3(\widetilde{I})$ . At this point, one introduces the steepness condition required for the application of Nekhoroshev theory by imposing that the quadratic form  $\widetilde{Q}$  is *sign definite* (a weaker condition is considered in [24]).

In a neighborhood of the considered torus, the problem is now reduced to the study of perturbations of a nonlinear *convex* integrable Hamiltonian and an exponential estimate of stability can be derived. These reasonings were stated in ([62], 2.2) and more specifically studied in [46,48].

A remarkable result of Morbidelli and Giorgilli [56] clearly shows that the two previous results which come respectively from Hamiltonian perturbation theory and Nekhoroshev's theorem are *independent* and can be *superimposed*. Indeed, in the case of a strongly nonresonant invariant torus (especially for a KAM torus) which admits *moreover* a sign definite torsion, one can state results

of stability over *superexponentially* long times. The proof starts with a Birkhoff's normal form like (10) but with an exponentially small remainder. Then, Nekhoroshev's theorem is applied with a perturbation which is already exponentially small, hence we obtain a superexponential time of stability. More specifically, provided that the Birkhoff normal form is quasi-convex, results of stability over times of the order of  $\exp(\exp(cR^{-1/\tau}))$  can be ensured for the solutions with an initial condition in a ball of radius  $R$  small enough around the invariant torus.

Actually, the previous results were extended [46,48] for an *elliptic equilibrium* point or a lower dimensional torus in a Hamiltonian system except for an annoying problem of singularity in the action-angle transformation. This problem does not allow one to prove a stability theorem for a complete neighborhood of an elliptic equilibrium point. The corresponding theorems for *all initial data* were obtained in [28,65,72] where Nekhoroshev's estimates were established without action-angle variables.

Finally, the relationship between KAM and Nekhoroshev theory was considered in [32] and [23] where the existence of a sequence of nested domains in phase space which converge to the KAM set of invariant tori for the perturbed system and over which stability estimates are valid and growth occurs in a superexponential way was proven. Especially, on the initial domain we recover the usual statement of Nekhoroshev.

## Applications

With the remaining space, we only give glimpses of application of the previous theorems in physics and astronomy. We mainly quote surveys in the sequel and these references do not form at all a complete list.

### The Case of a Constant Frequency Integrable System

The question of stability of a perturbed single frequency system corresponds to the problem of preservation of the *adiabatic invariants* which appears in numerous physical problems [2,59].

Especially, for the applications in plasma physics, we can mention the beautiful survey of Northrop [69]. The problem of charge trapping by strong nonuniform magnetic fields (Van Allen belts, magnetic bottles) can also be tackled by means of the adiabatic invariant theory [15].

The same question arises in connection with the problem of *energy equipartition* in large systems of Fermi–Pasta–Ulam type and the conjecture of Boltzmann and Jeans [8,30].

The problem of existence of an approximate first integral over a long time but this time for a quasi integrable

symplectic mapping appears to study the effective stability of the billiard flow near the boundary of a strictly convex domain in  $\mathbb{R}^n$  [35].

For several degrees of freedom, results of stability over very long times for a quasi integrable Hamiltonian system were first obtained by Littlewood [45] about *triangular Lagrangian equilibria* in the three bodies problem. One can reduce this question to the study of a perturbed strongly nonresonant constant frequency system [31].

For effective computations, it is much more efficient to make a numerical summation at the smallest term instead of plugging the data into an abstract theorem which usually gives poor estimates. Following this scheme, Giorgilli and different coauthors have obtained effective stability results in celestial mechanics with *realistic* physical parameters (see [8,27,30,33,74]).

The same situation of a perturbed strongly nonresonant constant frequency system appears in the study of stability of symplectic numerical integrators [14,41,55].

In the realm of PDE's, results of stability around *finite dimensional nonresonant tori* can be proved ([4,11,18,44] and ▶ [Perturbation Theory for PDEs](#) for surveys).

The remarkable paper of Bambusi and Grebert [6] gives an extension of the previous results for *infinite dimensional* nonresonant tori.

## Application of the Global Nekhoroshev Theory

Here, we look for global results of stability for perturbed nonlinear integrable Hamiltonian systems.

This situation appears in celestial mechanics where the unperturbed system is often *properly degenerate*, namely the number of constants of motion exceeds the number of degrees of freedom. This is the case for the Kepler problem which yields the integrable part of the planetary  $n$ -bodies problem (i. e.: the approximation of the  $n$ -bodies problem corresponding to the motion of the planets in the solar system). A study of this system in the light of Nekhoroshev's theory was given in [64], suggesting a modification of the original statement of Nekhoroshev by considering the proper degeneracies of this system.

The question of stability in celestial mechanics was also considered for the asteroid belt [57] where additional degeneracies and resonances appear (see also [40]).

We have seen that Nekhoroshev's theory also allows one to study stability around an elliptic equilibrium point in a Hamiltonian system with a sign definite torsion. But the most famous example of such an equilibrium in astronomy, namely the stability of an asteroid located at the top of an equilateral triangle with the Sun and Jupiter (the Lagrangian points L4, L5) cannot be tackled with the pre-

vious theorems in the convex case. It can be shown that with the actual masses of the Sun and Jupiter, the problem of stability of the Lagrangian points could be reduced to a study of a small perturbation of an integrable Hamiltonian of three degrees of freedom whose 3-jet satisfies a condition which implies steepness [10]. It allows one to prove a confinement over very long time intervals for asteroids located close enough to the Lagrangian points L4 or L5.

Other applications of Nekhoroshev's stability at an elliptic equilibrium point are the stability of the Riemann Ellipsoid [29], the fast rotation of a rigid body [9].

A Nekhoroshev-like theory has also been developed for beam dynamics [25,26].

From a numerical point of view, a new spectral formulation of the Nekhoroshev theorem has been introduced [38]. This allows one to recognize whether or not the motion in a quasi-integrable Hamiltonian system is in a Nekhoroshev regime (i. e. the action coordinates are eventually subject to an exponentially slow drift) by looking at the Fourier spectrum of the solutions.

In a geometric setting, an extension of Nekhoroshev's results for perturbations of convex, *noncommutatively* integrable Hamiltonian systems has been given in [16].

Finally, no general results like the Nekhoroshev theorem are known yet for large Hamiltonian systems or for Hamiltonian PDE's seen as infinite dimensional Hamiltonian systems. But a number of quasi-Nekhoroshev theorems for special systems of this type have been proved mostly by Bourgain and Bambusi (see [5] for large systems and [3,7,18] for PDE's).

## Future Directions

The analyticity of the studied systems is only needed for the construction of the normal forms up to an exponentially high remainder. On the other hand, the steepness condition is generic for Hamiltonians of *finite* but sufficiently high smoothness [68]. It would be natural to prove the analogous stability theorems in the case of smooth functions which would give stability over *polynomially* long times.

Another question is the extension of Lochak's mechanism of stabilization around resonances for *non* quasi-convex integrable Hamiltonians. These improved exponents are at the basis of Nekhoroshev-type results which are obtained in large systems [5] or in PDE [3,7,18] hence their generalization would be important.

The global stability results considered so far are obviously valid only if one takes into account the most pessimistic estimations in the whole phase space. On the other

hand, we have just seen that these results can be improved locally. It would be relevant to make a study of the *average* exponent of stability. It could be a *space* average of these exponents by considering all initial conditions or, in certain examples, the *time* average of these exponents by taking into account the variations of the speed of drift of the actions under Arnold's diffusion.

The applications of Nekhoroshev's theory in Astronomy is an active field either analytically or in a numerical way [36].

The relevance of Nekhoroshev's estimates for statistical mechanics and thermodynamics is an important question which is tackled with the problem of energy equipartition in the Fermi-Pasta-Ulam (FPU) model (see [19,20] and the references therein).

Finally, a partial generalization of KAM theory to PDEs has been carried out during the last twenty years by many high level mathematicians, but the theory developed up to now only allows one to show that *finite* dimensional invariant tori persist under perturbation. Thus, most of the initial data are outside invariant tori. It is therefore clear that it would be very important to understand the behavior of solutions starting outside the tori. This is also related to the problem of estimating the time of existence of the solutions of hyperbolic PDEs in compact domains, a problem that is one of the most important open questions in relation to hyperbolic PDEs. Up to now Nekhoroshev's theorem has been generalized to PDE's only in order to deal with small amplitude solutions but the obtention of really global results of stability would be a challenge (Kuksin has announced these kinds of theorems for the KdV equation, see ► [Perturbation Theory for PDEs](#)).

## Appendix: An Example of Divergence Without Small Denominators

We would like to develop the following example of Neishtadt [60] where the phenomenons of divergence without small denominators and summation up to an exponentially small remainder appear in its simplest setting (see also [76] for another example).

Consider the quasi integrable system governed by the Hamiltonian

$$\mathcal{H}(I_1, I_2; \theta_1, \theta_2) = I_1 - \varepsilon[I_2 - \cos(\theta_1)f(\theta_2)]$$

defined over  $\mathbb{R}^2 \times \mathbb{T}^2$ , (11)

with

$$f(\theta) = \sum_{m \geq 1} \frac{\alpha_m}{m} \cos(m\theta),$$

where  $\alpha_m = e^{-\alpha m}$  for  $0 < \alpha \leq 1$ , this last choice corresponds to the exponential decay of the Fourier coefficients for a holomorphic function.

Indeed,  $f'(\theta) = \text{Im}(g(-\alpha + i\theta))$  where  $g(z) = (\exp(z) - 1)^{-1}$  and the complex pole of  $f$  which is closest to the real axis is located at  $-i\alpha$ .

Conversely, all real function which admits a holomorphic extension over a complex strip of width  $\alpha$  (i. e.:  $|\text{Im}(z)| \leq \alpha$ ) has Fourier coefficients bounded by  $(C\alpha_m)_{m \in \mathbb{N}^*}$  for some constant  $C > 0$ .

As in the Sect. “[Hamiltonian Perturbation Theory](#)”, it is possible to eliminate formally completely the fast angle  $\theta_1$  in the perturbation *without the occurrence* of any small denominators.

Indeed, one can consider a normalizing transformation generated by  $X(\theta_1, \theta_2) = \sum_{n \geq 1} \varepsilon^n X_n(\theta_1, \theta_2)$  where the functions  $X_n$  satisfy the homological equations:

$$\begin{aligned} \partial_{\theta_1} X_1(\theta_1, \theta_2) &= \cos(\theta_1) f(\theta_2) \text{ and} \\ \partial_{\theta_1} X_n(\theta_1, \theta_2) &= \partial_{\theta_2} X_{n-1}(\theta_1, \theta_2). \end{aligned}$$

The solutions can be written  $X_n(\theta_1, \theta_2) = \cos(\theta_1 - n\frac{\pi}{2}) f^{(n-1)}(\theta_2)$ , hence:

$$\begin{aligned} X(\theta_1, \theta_2) &= \sum_{n \in \mathbb{N}^*} \varepsilon^n \sum_{m \in \mathbb{N}^*} \frac{\alpha_m}{m^2} m^n \\ &\quad \cdot \cos\left(\theta_1 - n\frac{\pi}{2}\right) \sin\left(m\theta_2 + n\frac{\pi}{2}\right) \end{aligned}$$

and, for instance,

$$X\left(\frac{\pi}{2}, 0\right) = \sum_{m \in \mathbb{N}^*} \sum_{k \in \mathbb{N}} \frac{e^{-\alpha m}}{m^2} (\varepsilon m)^{2k+1}$$

which is divergent for all  $\varepsilon > 0$  since  $\varepsilon \geq 1/m$  for  $m$  large enough.

We see that divergence comes from coefficients arising with successive differentiations.

On the other hand, if one considers the transformation generated by the truncated Hamiltonian  $\sum_{n \geq 1}^N \varepsilon^n X_n(\theta_1, \theta_2)$  then the transformed Hamiltonian becomes

$$\mathcal{H}(I_1, I_2; \theta_1, \theta_2) = I_1 - \varepsilon I_2 + \varepsilon^{N+1} \partial_{\theta_2} X_N(\theta_1, \theta_2),$$

where  $\partial_{\theta_2} X_N(\theta_1, \theta_2) = \cos(\theta_1 - N\frac{\pi}{2}) f^{(N)}(\theta_2)$ .

Especially, with  $g(z) = 1/\exp(z) - 1$ , we have:

$$\begin{aligned} \partial_{\theta_2} X_{2N}(\theta_1, \theta_2) &= -\cos(\theta_1) \text{Re}\left(g^{(2N-1)}(-\alpha + i\theta_2)\right) \\ &\quad \text{for } N > 0, \end{aligned}$$

$$\begin{aligned} \partial_{\theta_2} X_{2N+1}(\theta_1, \theta_2) &= \sin(\theta_1) \text{Im}\left(g^{(2N)}(-\alpha + i\theta_2)\right) \\ &\quad \text{for } N \geq 0. \end{aligned}$$

Now, around the real axis, the main term in the asymptotic expansion of  $g^{(n)}(z)$  as  $n$  goes to infinity is given by the derivative of the polar term  $1/z$  in the Laurent expansion of  $g$  at 0.

Indeed, we consider the function  $h(z) = g(z) - 1/z$  which is real analytic and admits an analyticity width of size  $2\pi$  around the real axis. Hence, Cauchy estimates ensure that for  $n$  large enough and  $z$  in the strip of width  $\pi$  around the real axis, the derivative  $h^{(n)}(z)$  becomes negligible with respect to the derivative of  $1/z$ .

Consequently,  $g^{(n)}(z)$  is *equivalent* to  $(-1)^n n!/z^{n+1}$  as  $n$  goes to infinity for  $z$  in the strip of width  $\pi$  around the real axis.

Hence, the remainder  $\partial_{\theta_2} X_N(\theta_1, \theta_2)$  has a size of order  $(N-1)!/\alpha^N$  for  $N$  large. This latter estimate *cannot be improved*, since  $\partial_{\theta_2} X_{2N}(\pi, 0)$  and  $\partial_{\theta_2} X_{2N+1}(\pi/2, \alpha \tan(\pi/4N + 2))$  admit the same size of order  $(N-1)!/\alpha^N$  for  $N$  large.

Now, we can make a “summation at the smallest term” as in Sect. “[The Case of a Single Frequency System](#)” to obtain an optimal normalization with an *exponentially small* remainder.

## Bibliography

### Primary Literature

1. Arnold VI (1964) Instability of dynamical systems with several degrees of freedom. *Sov Math Dokl* 5:581–585
2. Arnold VI, Kozlov VV, Neishtadt AI (2006) Mathematical aspects of classical and celestial mechanics, 3rd revised edn. In: *Encyclopaedia of Mathematical Sciences* 3. Dynamical Systems 3. Springer, New York
3. Bambusi D (1999) Nekhoroshev theorem for small amplitude solutions in nonlinear Schrödinger equations. *Math Z* 230(2):345–387
4. Bambusi D (1999) On long time stability in Hamiltonian perturbations of non-resonant linear PDEs. *Nonlinearity* 12(4):823–850
5. Bambusi D, Giorgilli A (1993) Exponential stability of states close to resonance in infinite-dimensional Hamiltonian systems. *J Stat Phys* 71(3–4):569–606
6. Bambusi D, Grébert B (2006) Birkhoff normal form for partial differential equations with tame modulus. *Duke Math J* 135(3):507–567
7. Bambusi D, Nekhoroshev NN (2002) Long time stability in perturbations of completely resonant PDE’s. *Acta Appl Math* 70(1–3):1–22
8. Benettin G (2005) Physical applications of Nekhoroshev theorem and exponential estimates. In: Giorgilli A (ed) *Cetraro (2000) Hamiltonian Dynamics, Theory and Applications*. Springer, New York
9. Benettin G, Fasso F (1996) Fast rotations of the rigid body: A study by Hamiltonian perturbation theory, I. *Nonlinearity* 9(1):137–186

10. Benettin G, Fasso F, Guzzo M (1998) Nekhoroshev stability of L4 and L5 in the spatial restricted three body problem. *Regul Chaotic Dyn* 3(3):56–72
11. Benettin G, Fröhlich J, Giorgilli A (1988) A Nekhoroshev-type theorem for Hamiltonian systems with infinitely many degrees of freedom. *Commun Math Phys* 119(1):95–108
12. Benettin G, Galgani L, Giorgilli A (1985) A proof of Nekhoroshev's theorem for the stability times in nearly-integrable Hamiltonian systems. *Celest Mech* 37:1–25
13. Benettin G, Gallavotti G (1986) Stability of motions near resonances in quasi-integrable Hamiltonian systems. *J Stat Phys* 44(3–4):293–338
14. Benettin G, Giorgilli A (1994) On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms. *J Stat Phys* 74(5–6):1117–1143
15. Benettin G, Sempio P (1994) Adiabatic invariants and trapping of a point charge in a strong non-uniform magnetic field. *Nonlinearity* 7(1):281–303
16. Blaom AD (2001) A geometric setting for Hamiltonian perturbation theory. *Mem Am Math Soc* 727(xviii):112
17. Bogolyubov NN, Mitropol'skij YA (1958) *Asymptotic Methods in the Theory of Nonlinear Oscillations*, 2nd edn. Nauka, Moscow. Engl. Transl. (1961) Gordon and Breach, New York
18. Bourgain J (2004) Remarks on stability and diffusion in high-dimensional Hamiltonian systems and partial differential equations. *Ergod Theory Dyn Syst* 24(5):1331–1357
19. Carati A, Galgani L, Giorgilli A, Ponno A (2002) The Fermi–Pasta–Ulam Problem. *Nuovo Cimento B* 117:1017–1026
20. Carati A, Galgani L, Giorgilli A (2006) *Dynamical Systems and Thermodynamics*. In: Françoise JP, Naber GL, Tsun TS (eds) *Encyclopedia of mathematical physics*. Elsevier, Amsterdam
21. 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(1):31–58
22. Chirikov BV (1979) A universal instability in many dimensional oscillator systems. *Phys Rep* 52:263–279
23. Delshams A, Gutierrez P (1996) Effective Stability and KAM Theory. *J Diff Eq* 128(2):415–490
24. Dullin H, Fassò F (2004) An algorithm for detecting directional quasi-convexity. *BIT* 44(3):571–584
25. Dumas HS (1993) A Nekhoroshev-like theory of classical particle channeling in perfect crystals. In: Jones CKRT et al (ed) *Dynamics reported. Expositions in dynamical systems, new series*, vol 2. Springer, Berlin
26. Dumas HS (2005) Mathematical theories of classical particle channeling in perfect crystals. *Nucl Inst Meth Phys Res Sect B (Beam Interactions with Materials and Atoms)* 234(1–2):3–13
27. Efthymiopoulos C, Giorgilli A, Contopoulos G (2004) Nonconvergence of formal integrals. II: Improved estimates for the optimal order of truncation. *J Phys A (Math Gen)* 37(45):10831–10858
28. Fassò F, Guzzo M, Benettin G (1998) Nekhoroshev-stability of elliptic equilibria of Hamiltonian systems. *Commun Math Phys* 197(2):347–360
29. Fassò F, Lewis D (2001) Stability properties of the Riemann ellipsoids. *Arch Ration Mech Anal* 158(4):259–292
30. Giorgilli A (1998) On the problem of stability for near to integrable Hamiltonian systems. In: *Proceedings of the International Congress of Mathematicians Berlin 1998. Documenta Mathematica III*:143–152
31. 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 bodies problem. *J Diff Equa* 77:167–198
32. Giorgilli A, Morbidelli A (1997) Invariant KAM tori and global stability for Hamiltonian systems. *Z Angew Math Phys* 48(1):102–134
33. Giorgilli A, Skokos C (1997) On the stability of the Trojan asteroids. *Astron Astroph* 317:254–261
34. Giorgilli A, Zehnder E (1992) Exponential stability for time dependent potential. *ZAMP* 43:827–855
35. Gramchev T, Popov G (1995) Nekhoroshev type estimates for billiard ball maps. *Ann Inst Fourier* 45(3):859–895
36. Guzzo M (2003) Nekhoroshev stability of asteroids. Triennial report 2000–2003 of Commission 7-Celestial Mechanics and dynamical Astronomy of the IAU; Reports on Astronomy, 1999–2002. Transactions of the International Astronomical Union, vol XXVA, Astronomical Society of the Pacific
37. Guzzo M (2004) A direct proof of the Nekhoroshev theorem for nearly integrable symplectic maps. *Ann Henri Poincaré* 5(6):1013–1039
38. Guzzo M, Benettin G (2001) A spectral formulation of the Nekhoroshev theorem and its relevance for numerical and experimental data analysis. *Discret Contin Dyn Syst Ser B* 1(1):1–28
39. Guzzo M, Lega E, Froeschlé C (2006) Diffusion and stability in perturbed non-convex integrable systems. *Nonlinearity* 19(5):1049–1067
40. Guzzo M, Morbidelli A (1997) Construction of a Nekhoroshev-like result for the asteroid belt dynamical system *Celest. Mech Dyn Astron* 66:255–292
41. Hairer E, Lubich C, Wanner G (2006) *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations*, 2nd edn. Springer Series in Computational Mathematics, vol 31. Springer, New York
42. Ilyashenko IS (1986) A steepness test for analytic functions. *Russ Math Surv* 41:229–230
43. Kuksin S, Pöschel J (1994) On the inclusion of analytic symplectic maps in analytic Hamiltonian flows and its applications. In: Kuksin S, Lazutkin VF, Pöschel J (eds) *Proceedings of the 1991 Euler Institute Conference on Dynamical Systems. Prog NonLin Diff Equ App* (12) Birkhäuser, Basel
44. Kuksin SB (2006) *Hamiltonian PDEs (with an appendix by Dario Bambusi)*. In: Hasselblatt B (ed) *Handbook of dynamical systems*, vol 1B. Elsevier, Amsterdam
45. Littlewood JE (1959) On the equilateral configuration in the restricted problem of the three bodies. *Proc London Math Soc* 9(3):343–372
46. Lochak P (1992) Canonical perturbation theory via simultaneous approximation. *Russ Math Surv* 47:57–133
47. Lochak P (1993) Hamiltonian perturbation theory: Periodic orbits, resonances and intermittency. *Nonlinearity* 6(6):885–904
48. Lochak P (1995) Stability of Hamiltonian systems over exponentially long times: the near-linear case. In: Dumas HS, Meyer K, Schmidt D (eds) *Hamiltonian dynamical systems – History, theory and applications. IMA conference proceedings series*, vol 63. Springer, New York, pp 221–229
49. Lochak P, Meunier C (1988) Multiphase averaging methods for Hamiltonian systems. *Appl Math Sci Series*, vol 72. Springer, New York



50. Lochak P, Neishtadt AI (1992) Estimates of stability time for nearly integrable systems with a quasiconvex Hamiltonian. *Chaos* 2(4):495–499
51. Lochak P, Neishtadt AI, Niederman L (1994) Stability of nearly integrable convex Hamiltonian systems over exponentially long times. In: Kuksin S, Lazutkin VF, Pöschel J (eds) *Proceedings of the 1991 Euler Institute Conference on Dynamical Systems. Prog NonLin Diff Equ App* (12). Birkhäuser, Basel
52. Marco JP, Lochak P (2005) Diffusion times and stability exponents for nearly integrable analytic systems. *Central Eur J Math* 3(3):342–397
53. Marco JP, Sauzin D (2003) Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian Systems. *Publ Math Inst Hautes Etudes Sci* 96:199–275
54. Meyer KR, Hall GR (1992) *Introduction to Hamiltonian dynamical systems and the N-Body problem*. Applied Mathematical Sciences, vol 90. Springer, New York
55. Moan PC (2004) On the KAM and Nekhoroshev theorems for symplectic integrators and implications for error growth. *Nonlinearity* 17(1):67–83
56. Morbidelli A, Giorgilli A (1995) Superexponential stability of KAM tori. *J Stat Phys* 78:1607–1617
57. Morbidelli A, Guzzo M (1997) The Nekhoroshev theorem and the asteroid belt dynamical system. *Celest Mech Dyn Astron* 65(1–2):107–136
58. Moser J (1955) Stabilitätsverhalten Kanonischer Differentialgleichungssysteme. *Nachr Akad Wiss Göttingen, Math Phys Kl IIa* 6:87–120
59. Neishtadt AI (1981) On the accuracy of conservation of the adiabatic invariant. *Prikl Mat Mekh* 45:80–87. Translated in *J Appl Math Mech* 45:58–63
60. Neishtadt AI (1984) The separation of motions in systems with rapidly rotating phase. *J Appl Math Mech* 48:133–139
61. Nekhoroshev NN (1973) Stable lower estimates for smooth mappings and for gradients of smooth functions. *Math. USSR Sb* 19(3):425–467
62. Nekhoroshev NN (1977) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. *Russ Math Surv* 32:1–65
63. Nekhoroshev NN (1979) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems 2. *Trudy Sem Petrovs* 5:5–50. Translated In: Oleinik OA (ed) *Topics in Modern Mathematics. Petrovskii Semin*, vol 5. Consultant Bureau, New York
64. Niederman L (1996) Stability over exponentially long times in the planetary problem. *Nonlinearity* 9(6):1703–1751
65. Niederman L (1998) Nonlinear stability around an elliptic equilibrium point in an Hamiltonian system. *Nonlinearity* 11:1465–1479
66. Niederman L (2004) Exponential stability for small perturbations of steep integrable Hamiltonian systems. *Erg Theor Dyn Syst* 24(2):593–608
67. Niederman L (2006) Hamiltonian stability and subanalytic geometry. *Ann Inst Fourier* 56(3):795–813
68. Niederman L (2007) Prevalence of exponential stability among nearly integrable Hamiltonian systems. *Erg Theor Dyn Syst* 27(3):905–928
69. Northrop TG (1963) *The adiabatic motion of charged particles*. Interscience Publishers, New York
70. Poincaré H (1892) *Méthodes Nouvelles de la Mécanique Céleste*, vol 4. Blanchard, Paris
71. Pöschel J (1993) Nekhoroshev estimates for quasi-convex Hamiltonian systems. *Math Z* 213:187–217
72. Pöschel J (1999) On Nekhoroshev's estimate at an elliptic equilibrium. *Int Math Res Not* 1999(4):203–215
73. Ramis JP, Schäfke R (1996) Gevrey separation of fast and slow variables. *Nonlinearity* 9(2):353–384
74. Steichen D, Giorgilli A (1998) Long time stability for the main problem of artificial satellites. *Cel Mech* 69:317–330
75. Treschev DV (1994) Continuous averaging in Hamiltonian systems. In: Kuksin S, Lazutkin VF, Pöschel J (eds) *Proceedings of the 1991 Euler Institute Conference on Dynamical Systems. Prog NonLin Diff Equ App* (12). Birkhäuser, Basel
76. Valdinoci E (2000) Estimates for non-resonant normal forms in Hamiltonian perturbation theory. *J Stat Phys* 101(3–4):905–919

### Books and Reviews

- Arnold VI (1983) *Geometrical methods in the theory of ordinary differential equations*. Transl. from the Russian by Joseph Szeecs, Mark Levi (ed) *Grundlehren der Mathematischen Wissenschaften* 250. Springer, New York
- Arnold VI, Kozlov VV, Neishtadt AI (2006) *Mathematical aspects of classical and celestial mechanics*, 3rd revised edn. *Encyclopaedia of Mathematical Sciences* 3. Dynamical Systems 3. Springer, New York
- Benettin G (2005) Physical applications of Nekhoroshev theorem and exponential estimates. In: Giorgilli A (ed) *Cetraro (2000) Hamiltonian Dynamics, Theory and Applications*. Springer, New York
- Giorgilli A (2003) Exponential stability of Hamiltonian systems. *Dynamical systems, Part I. Hamiltonian systems and celestial mechanics*. Selected papers from the Research Trimester held in Pisa, Italy, February 4–April 26, 2002. Pisa: Scuola Normale Superiore. *Pubblicazioni del Centro di Ricerca Matematica Ennio de Giorgi*. Proceedings, 87–198
- Sanders JA, Verhulst F, Murdock J (2007) *Averaging methods in nonlinear dynamical systems*, 2nd edn. *Applied Mathematical Sciences* 59. Springer, New York

---

## Network Analysis, Longitudinal Methods of

TOM A. B. SNIJDERS

University of Oxford, Oxford, United Kingdom

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Stochastic Models for Network Dynamics](#)

[Statistical Estimation and Testing](#)

[Example: Dynamics of Adolescent Friendship](#)

## Models for the Co-evolution of Networks and Behavior

### Example: Co-evolution of Adolescent Friendship and Alcohol Use

#### Extensions

#### Future Directions

#### Bibliography

## Glossary

**Actors** The social actors who are represented by the nodes of the network, and indicated by a label denoted  $i$  or  $j$  in the set  $1, \dots, n$ .

**Behavior** An umbrella term for changing characteristics of actors, considered as components of the outcome of the stochastic system: e. g., behavioral tendencies or attitudes of human actors, performance, etc. Each behavior variable  $Z_h$  is assumed to be measured on an ordinal discrete scale with values  $1, 2, \dots, M_h$  for some  $M_h \geq 2$ . The value of behavior variable  $Z_h$  for actor  $i$  is denoted  $Z_{ih}$ .

**Change determination process** The stochastic model defining the probability distribution of changes, conditional on the event that there is an opportunity for change.

**Change opportunity model** The stochastic process defining the moments where tie indicators can change. This can be either *tie-based*, meaning that an ordered pair of actors ( $i, j$ ) is chosen and the possibility arises that the tie variable from  $i$  to  $j$  is changed; or *actor-based*, meaning that an actor  $i$  is chosen and the possibility arises that one of the outgoing tie variables from actor  $i$  is changed.

**Covariates** Variables which can depend on the actors (actor covariates) or on pairs of actors (dyadic covariates), and which are considered to be deterministic, or determined outside of the 'stochastic system' under consideration.

**Effects** Components of the objective function.

**Influence** The phenomenon that change probabilities for actors' behavior depend on the network positions of the actors, usually in combination with the current behavior of the other actors.

**Markov chain** A stochastic process where the probability distribution of future states, given the present state, does not depend on past states.

**Method of moments** A general method of statistical estimation, where the parameters are estimated in such a way that expected values of a vector of selected statistics are equal to their observed values.

**Network** A simple directed graph representing a relation on the set of actors with binary tie indicators  $X_{ij}$  which can be regarded as a state which can change, but will normally change slowly.

**Objective function** Usually denoted by  $f_i$ ; the informal description is that this is a measure of how attractive it is to go from an old to a new state. More formally, when there is an opportunity for change, the probability of the change is assumed to be proportional to the exponential transform of the objective function.

The objective function has a similar role as the linear predictor in generalized linear models in statistics, and is specified here as a linear combination of *effects*.

**Rate function** Usually denoted by  $\lambda$ , the expected number of opportunities for change per unit of time.

**Selection** The phenomenon that change probabilities for network ties depend on the behavior of one or both of the two actors involved.

**Tie indicator** A variable  $X_{ij}$  indicating by the value  $X_{ij} = 1$  that there is a tie  $i \rightarrow j$ , and by the value  $0$  that there is no such tie. Also called tie variables.

## Definition of the Subject

Social networks represent the patterns of ties between social actors. To analyze empirically the mechanisms that determine creation and termination of ties, especially if several mechanisms that may be complementary are studied simultaneously, statistical methods are needed. This chapter is aimed at the case that network panel data are available to the researcher, and treats recently developed statistical models for such data, with corresponding estimation methods. To represent the feedback processes inherent in network dynamics, it is helpful to regard such panel data as momentary observations on a continuous-time stochastic process on the space of directed graphs. Tie-oriented and actor-oriented stochastic models are presented, which can reflect endogenous network dynamics as well as effects of exogenous variables. These models can be regarded as agent-based models, and they can be implemented as computer simulation models. To estimate the parameters of the model, stochastic approximation methods can be used. Social networks are especially interesting because they are important influences on individual behavior – and in turn the network ties are influenced by individual behavior. This two-way influence can be represented by models for the co-evolution of networks and changeable actor attributes. Such models, and statistical methods to analyze panel data on networks and behavior, are also treated. An extensive example is discussed about friendship among teenagers in a school setting.

## Introduction

When we think of social networks, it is quite natural to think of them as being dynamic. Ties are established, gain in strength, they can blossom and decay, and they may wither or be terminated with a bang. This applies to all kinds of relations – friendship or collaboration between humans, joint ventures between companies, bilateral agreements between countries, and so on. (Note that all these are examples of ‘positive’ ties; ‘negative’ ties are not dealt with in this chapter). It is also natural to think of such changes as being dependent on characteristics of the nodes – inclinations and abilities of humans, resources of companies, locations and capacities of countries – and on characteristics of pairs of nodes such as similarity or spatial proximity, as well as on the existing network structure – reciprocation of friendships, transitive closure of friendships (which is the case when friends of friends become friends), group formation of companies or countries. Finally, we are not surprised when such network changes have repercussions – friendships are often thought to have good or bad influences on the individuals concerned, agreements between companies and between countries will have consequences for the performance of the companies and for how the countries fare. Indeed, in the cases of the companies and countries, the links often are created with the purpose of having beneficial consequences for the companies or countries, respectively.

This sets the stage for this chapter, which is concerned with inferential data analysis for network dynamics. The focus is on social networks, and the nodes, or vertices, in the network will be referred to as (social) actors. Data analysis means that methods will be presented for analyzing empirical data on network change. Inference means that the aim is to have methods for testing hypotheses about mechanisms that may drive the network dynamics, and for estimating parameters that figure in such mechanisms. As usual in statistical inference, the ‘mechanisms’ will be expressed as probability models, also called stochastic models. For hypothesis testing it must be possible to play off one theory against another, which means that when we have theories, or mechanisms,  $T_1$  and  $T_2$ , that are not contradictory but could occur together, we need models which can express both mechanisms simultaneously and reflect each mechanism in a set of parameters. The mechanism operates if some of its parameters are nonzero. Then we can test the null hypothesis that  $T_1$  operates but not  $T_2$ , against the alternative hypothesis that both mechanisms operate; in other words, we can test for  $T_2$  while controlling for  $T_1$ . This requires flexibility of the stochastic mod-

els being used: what was stated above implies that it must be possible to specify the models in diverse ways, e. g., with some parameters for effects of actor characteristics and others for effects of various aspects of the existing network structure, and these parameters being estimable from observed data. For data consisting of independent observations the elaboration of these principles of statistical inference in the linear regression model and its generalizations is well known. Network data have the complicating feature that they are not composed of independent observations: the occurrence, creation, or termination of one tie is highly dependent on the existence of other ties. Therefore more complex statistical models are needed, giving an adequate representation of the mutual dependence between the existence of various ties between the actors in the network. But readers who are acquainted with generalized linear statistical models will see that, although the models treated in this chapter have this greater complexity, many elements known from generalized linear models do play a role.

Social network analysis is concerned with diverse types of networks which can be represented by diverse data structures: simple graphs and their generalizations. In line with the current state of the statistical methodology for network dynamics, this chapter is restricted to data structures where the changing network is a changing simple directed graph, where the arcs represent social ties that can be regarded as *states* rather than events. This means that, although changeable, the ties have inertia, a tendency to endure. Friendship between humans and agreements between companies are examples of states; conversations and momentary transactions are events. For networks where ties are states, the dependence between ties can be represented by assuming that changes in the network are dependent upon the existing network structure. In mathematical terms, this is to say that it is reasonable to assume that – given the available ‘independent’ or ‘explanatory’ variables – the network is a *Markov chain*. A Markov chain is a stochastic process where the probability distribution of future states, given the present state, does not depend on the past states. Such models were first proposed for network dynamics by [22] and elaborated by [59]. For some sociological applications outside the realm of networks, see [2,10]. For a network of events (e. g., the network of conversations ongoing at each given moment) the assumption of a Markov chain would be untenable. For a network of states (e. g., the network of joint ventures between companies) the Markov assumption is usually not totally realistic but can be used as a first approximation, and in many cases is the best assumption one can make given the limitation of the available data. This assumption

often can be made more plausible by using relevant explanatory variables.

In the start of this introduction, ties were portrayed as possibly gaining in strength or decaying. It would be attractive to reflect this by measuring the ties on an ordinal scale. This is not considered here: we are dealing with simple and not with valued directed graphs, and the *tie indicators*, which are the variables  $X_{ij}$  indicating how actor  $i$  is tied to actor  $j$ , are constrained to having the values 0 or 1, indicating, respectively, absence and presence of the tie  $i \rightarrow j$ . The restriction to binary tie variables is in line with traditional network analysis, but extensions to valued ties are important and are the subject of current work. The Markov assumption will often be more reasonable for valued than for binary ties.

The assumption that we are dealing with a network of relational states must be reflected by the way in which the ties are measured. Measurement in social network analysis is a subject which tends to receive too little attention, and this chapter is no exception. In practice, ties will have to be measured in such a way that observing a tie is a good indicator for the relational state being investigated, such as friendship or collaboration. When the relation under study is a type of communication, which usually has an ephemeral nature, it will be necessary to aggregate the communication over a sufficiently long time interval, so that the resulting variable can be regarded as indicative of a relational state. For example, [25] aggregated email communication to the binary tie variable defined by at least one email being sent over a 60-day period. In other situations, shorter or longer periods may be relevant.

### Stochastic Models for Network Dynamics

This section presents stochastic models for use in statistical modeling. The inferential aspects (parameter estimation and testing of hypotheses) are treated in Section “Statistical Estimation and Testing”. These models were applied, e.g., to the testing of theories about dynamics of friendship networks [12,55,57], of trust networks in firms [56], of artistic prestige [13], and of ties between venture capital firms [8].

The network is represented by the node set  $\{1, \dots, n\}$  with tie variables  $x_{ij}$ , where  $x_{ij} = 1$  or 0 indicates whether the tie  $i \rightarrow j$  is present or absent. The tie variables are collected in the  $n \times n$  adjacency matrix  $x = (x_{ij})$ . Self-ties are excluded, so that  $x_{ii} = 0$  for all  $i$ . The concepts of network (directed graph) and matrix (its adjacency matrix) will be used interchangeably, depending on what is most convenient. In accordance with the usual notation in probability and statistics, random variables will be indicated by

capitals; and observations, or other non-random variables, by small letters. The ties are assumed to be outcomes of time-dependent random variables, denoted by  $X_{ij}(t)$  and collected in the time-dependent random matrix  $X(t)$ .

In addition to the network  $X(t)$ , which can be regarded as the dependent variable of the model, there can be other variables regarded as independent or explanatory variables in the sense that their values are not modeled but accepted as given, and they may influence the network. Such variables are called *covariates* and when depending on the actors they are denoted  $v_i$ , while if they depend on pairs of actors (*dyads*) the notation is  $w_{ij}$ . Examples are the age of actors (actor variable) and their spatial proximity (dyadic variable).

### Basic Model Definition

The following basic assumptions are made.

1. Time, denoted by  $t$ , is a continuous variable. This does not mean that it is assumed that observations are made continuously; in most practical cases, observations are made at a number (perhaps a small number) of discrete time moments. However, it is natural and mathematically convenient to assume that there is an underlying process  $X(t)$  (which may be observed only partially) which proceeds in continuous time.
2.  $X(t)$  is a Markov process.  
This means that the conditional distribution of future states depends on the past only as a function of the present. In other words, to predict the future it is sufficient to know the present state of the network, and knowledge of past states will not improve predictability. This assumption was discussed above. It can be expressed by saying that the network represents a *state*, and usually goes together with inertia, i.e., the tendency of ties to remain in existence unless something special happens.
3. At any given moment  $t$ , no more than one tie variable  $X_{ij}(t)$  can change.

This assumption, first proposed by [22], means that changes of ties are not directly coordinated, and ties are mutually dependent only because tie changes will depend on the current total configuration of ties. This is an important simplifying condition and excludes, for example, partner swapping and the coordinated formation of groups. Changes of several ties are decomposed as sequences of changes of single ties.

These assumptions still allow an extremely wide array of probability models, and further specification is necessary. In network change, two aspects can be distinguished: the frequency of tie change, which may depend on the ac-

tors involved; and the network structures that tend to be formed by the tie changes – the ‘direction’ of change. Examples of the former are that younger individuals might change their friendship ties more frequently than older individuals, or that more central actors might change their ties more frequently than peripheral actors. Examples of the latter are tendencies toward tie reciprocation, and toward transitive network closure. These two aspects will be represented by distinct components of the model and distinct parameters, which allow the inference about the one aspect to be relatively undisturbed by inference about the other aspect. The first component is the *change opportunity* process, the second the *change determination* model.

**Change Opportunity Process** Two specifications of the change opportunity process are given. They use the concept of a *Poisson process*, which is a stochastic process of events occurring at a certain rate  $\lambda$ , which means that the probability that an event occurs in the time interval from  $t$  to  $t + \epsilon$ , where  $\epsilon$  is a small positive number, is given (in the limit for  $\epsilon$  tending to 0) by  $\lambda\epsilon$ . One could say that the rate is the probability of occurrence per unit of time (in short time intervals).

Two specifications of the opportunity process are given here.

1. *Tie-based change opportunities*

For each tie variable  $X_{ij}$ , opportunities for change occur according to a Poisson process with rate  $\lambda_{ij}$ .

2. *Actor-based change opportunities*

For each actor  $i$ , opportunities to establish one new outgoing tie  $i \rightarrow j$ , or dissolve one existing tie  $i \rightarrow j$ , occur according to a Poisson process with rate  $\lambda_i$ .

The rates  $\lambda$  can be constant, or depend on covariates or functions of the current state of the network; if they are not constant, they are called *rate functions*. Tie-based rates  $\lambda_{ij}$ , for example, could depend on the proximity between actors or on their joint embeddedness such as the current number of common friends  $\sum_h X_{ih}(t) X_{jh}(t)$ . Actor-based rates  $\lambda_i$  could depend on actor variables or on positional variables such as actor  $i$ 's current outdegree  $\sum_j X_{ij}(t)$ .

Tie-based change opportunities were proposed by Robins and Pattison (personal communication), and correspond to the Gibbs sampling and Metropolis–Hastings procedures for simulating exponential random graph models, see [41]. Actor-based change opportunities were proposed by [46].

When an opportunity for change occurs, there is, in each opportunity model, a set of potential new networks that could be the result of the change. Denoting the current

network by  $x^0$ , for the tie-based opportunity model this set can be denoted by  $C_{ij}(x^0)$ . This is the set of the two possible matrices  $x$  where all elements other than  $x_{ij}$  are equal to those in the current matrix  $x^0$ , and where  $x_{ij}$  itself can be either 0 or 1. In the actor-based opportunity model actor  $i$ , when confronted with an opportunity for change, chooses one of his outgoing tie variables and changes this into its opposite value, changing 0 to 1 (creating a new tie) or changing 1 to 0 (terminating an existing tie). Therefore the set of potential new networks here is the set composed of  $x^0$  itself together with the  $n - 1$  matrices which are equal to  $x^0$  except for exactly one non-diagonal element in line  $i$  which is replaced by its opposite,  $x_{ij} = 1 - x_{ij}^0$ .

The set of new possible states is denoted in shorthand applicable to either case by  $C(x^0)$ . Since it is allowed that the current situation is continued, it always holds that  $x^0 \in C(x^0)$ .

**Change Determination Model** The choice of the new state of the network is dependent on what is called the *objective function*, which is a function  $f_i(x^0, x, v, w)$  depending on the current state of the network  $x^0$ , the potential new state  $x$ , the actor  $i$ , and the covariates summarized here as  $v$  (actor covariates) and  $w$  (dyadic covariates). The objective function can be interpreted informally as a measure of how attractive it is for actor  $i$  to change from state  $x^0$  to state  $x$ .

When actor  $i$  has the opportunity to change some outgoing tie variable  $X_{ij}$ , given that currently  $X(t) = x^0$ , the set of possible new states of the network is denoted  $C(x^0)$ . All  $x \in C(x^0)$  differ from  $x^0$  by at most one element  $x_{ij}$  for some  $j$ . When there is an opportunity for change from the current state  $x^0$ , the probabilities of the values of the next state  $x \in C(x^0)$  are proportional to  $\exp(f_i(x^0, x, v, w))$ .

The models can be summarized as follows. For the tie-oriented model, when an opportunity for change occurs, it refers to some pair  $(i, j)$ ; opportunities for changing  $X_{ij}$  occur at a rate  $\lambda_{ij}$  for each pair  $(i, j)$ . When such an opportunity occurs, the probability that  $x^0$  changes to the different state  $x$  is given by

$$\begin{aligned} & P\{X(t) \text{ changes to } x \mid (i, j) \text{ has a change opportunity at} \\ & \quad \text{time } t, X(t) = x^0\} \\ &= p_{ij}(x^0, x, v, w) \\ &= \frac{\exp(f_i(x^0, x, v, w))}{\exp(f_i(x^0, x^0, v, w)) + \exp(f_i(x^0, x, v, w))} \end{aligned} \quad (1)$$



where  $x$  and  $x^0$  are identical except for  $x_{ij} = 1 - x_{ij}^0$ .

For the actor-oriented model, opportunities for change occur for actors  $i$ . Opportunities for actor  $i$  to change one of the outgoing tie variables  $X_{ij}$  ( $j = 1, \dots, n$ ;  $j \neq i$ ) occur at a rate  $\lambda_i$ . The set of permitted new states, following on a given current state  $x^0$ , is  $C(x^0)$ . The probability that the new state is  $x$ , provided that  $x$  is permitted (i. e.,  $x \in C(x^0)$ ), is given by

$$\begin{aligned} & P\{X(t) \text{ changes to } x \mid i \text{ has a change opportunity at} \\ & \text{time } t, X(t) = x^0\} \\ &= p_i(x^0, x, v, w) \\ &= \frac{\exp(f_i(x^0, x, v, w))}{\sum_{x' \in C(x^0)} \exp(f_i(x^0, x', v, w))}. \end{aligned} \quad (2)$$

The two model components can be put together by giving the transition rate matrix, also called  $Q$ -matrix, of which the elements are defined by

$$q_{x^0, x} = \lim_{dt \downarrow 0} \frac{P\{X(t + dt) = x \mid X(t) = x^0\}}{dt} \quad (x \neq x^0)$$

(see textbooks on continuous-time Markov chains, such as [34]). Note that the assumptions imply that

$$q_{x^0, x} = 0 \text{ whenever } x_{ij} \neq x_{ij}^0 \text{ for more than one element } (i, j).$$

For digraphs  $x$  and  $x^0$  which differ from each other only in the element with index  $(i, j)$ , the elements of the  $Q$ -matrix are given for the tie-based opportunity process by

$$q_{x^0, x} = \lambda_{ij}(x^0, v, w) p_{ij}(x^0, x, v, w) \quad (3)$$

and for the actor-based opportunity process by

$$q_{x^0, x} = \lambda_i(x^0, v, w) p_i(x^0, x, v, w). \quad (4)$$

Tie-based models with constant change rates and objective functions defined as  $f(x, v, w)$  (not depending on the preceding state  $x^0$  or on the actor  $i$ ) can be regarded as Metropolis–Hastings dynamics (cf. [34]) for obtaining random draws from the digraph probability distribution with probability function

$$c \exp(f(x, v, w))$$

where  $c$  is a normalizing constant. In statistical mechanics  $f(x, v, w)$  then will be called a potential function, see [32]. When  $f(x, v, w)$  is a linear combination as in (6) below, the distribution is an exponential random graph model

for which the Metropolis–Hastings algorithm is treated in [41].

The actor-based opportunity model was proposed in [46] and, for a different data structure, in [45], as a stochastic actor-oriented model. In this model, the network dynamics is regarded as being driven by the social actors. Actors are assumed to control their outgoing ties, subject to inertia and the current network structure. This point of view is in accordance with the methodological approach of structural individualism [54,60], where actors are assumed to be purposeful and to behave subject to structural constraints. The purposes and constraints of the actors are summarized in the objective functions. One way to obtain the probabilities (2) is to assume that at each opportunity for change, actor  $i$  myopically optimizes the objective functions plus a random term, under the constraint that only one tie can change at a time. The myopia means that the actor only optimizes the state of the network that will be the immediate result of this change, without considering later network structures that might result in the future further ahead. The random term expresses otherwise unmodeled purposes and constraints. When the random terms have independent Gumbel distributions (the precise mathematical form of which does not matter for the present exposition), the choice probabilities are given by (2) (cf. [30,45,46]).

**Simulation** The Markov process defined above can be iterating by the following algorithm. The various steps in the algorithm can be derived using basic properties of Poisson processes and conditional probabilities (see [34]).

1. The process starts with a given time  $t$  and current state  $X(t) = x^0$ .
2. For the tie-based opportunity process define  $\lambda = \sum_{ij} \lambda_{ij}$ , and for the actor-based opportunity process  $\lambda = \sum_i \lambda_i$ . Let  $U$  be an independently drawn random number, uniformly distributed between 0 and 1, and let  $\Delta t = -\ln(U)/\lambda$ . Note that  $\Delta t$  has the exponential distribution with parameter  $\lambda$ . Change  $t$  into  $t + \Delta t$ .
- 3a. In the case of the tie-based opportunity process, choose a random pair  $(i, j)$  (with  $i \neq j$ ) with probabilities  $\lambda_{ij}/\lambda$ . With probability given by (1), change  $X_{ij}(t)$  into  $1 - x_{ij}^0$ .
- 3b. In the case of the actor-based opportunity process, choose a random actor  $i$  with probabilities  $\lambda_i/\lambda$ . To have a way for denoting the permitted new digraphs which are elements of  $C_i(x^0)$ , define by  $x^0(i \rightsquigarrow j)$  for  $j \neq i$  the digraph which is equal to  $x^0$  except

only that  $(x^0(i \rightsquigarrow j))_{ij} = 1 - x_{ij}^0$ ; define  $x^0(i \rightsquigarrow i) = x^0$ . Then choose a random  $j$  with probabilities

$$p_i(x^0, x^0(i \rightsquigarrow j), v, w) = \frac{\exp(f_i(x^0, x^0(i \rightsquigarrow j), v, w))}{\sum_{x' \in C(x^0)} \exp(f_i(x^0, x', v, w))}, \quad (5)$$

which is just the same as (2). If  $j \neq i$ , change  $X_{ij}(t)$  into  $1 - x_{ij}^0$ .

4. Go to step 1.

The stochastic process on the space of digraphs can be defined by this simulation algorithm just as well as by the  $Q$ -matrices (3) and (4), respectively.

### Specification of the Change Determination Process

The changes can be regarded for tie-based opportunities as determinations of new values for  $X_{ij}$  according to a binary logistic regression model (see [23]); and for actor-based opportunities, according to a multinomial logistic regression model (see [29]). In the further elaboration the parallel with logistic regression is followed because the objective function  $f_i$  is specified as a linear combination

$$f_i(x^0, x, v, w) = \sum_k \beta_k s_{ki}(x^0, x, v, w) \quad (6)$$

where the functions  $s_{ki}$  are so-called *effects* driving the network dynamics while the weights  $\beta_k$  are parameters indicating the force of these effects and which can be estimated from the data.

The specification of the model will be the choice of a limited set of such effects for use in (6). A list of some effects is the following. In the formulae, replacing an index by a + means that a sum is taken over this index. Many examples of effects do not depend on  $x^0$  or on  $v$  or  $w$ , and these arguments are then dropped from the notation.

**Outdegree effect**  $s_{1i}(x) = x_{i+} = \sum_j x_{ij}$  This effect models the tendency to have ties at all; this tendency will also be influenced by all other effects, and therefore the interpretation of its parameter is conditional on the further selection of effects included in the model. In many models this is the only effect to which those actors  $j$  contribute who have no reciprocal link to  $i$  nor any links with any others to whom  $i$  is linked. In such models, its weight  $\beta_1$  can be interpreted as the ‘value’ for actor  $i$  of a tie to such an other actor who is further completely isolated from  $i$ ’s personal network.

**Reciprocity effect**  $s_{2i}(x) = \sum_j x_{ij} x_{ji}$  This is the number of reciprocated ties for actor  $i$ . It models the tendency toward reciprocation of choices. Thus, a higher

value for its parameter  $\beta_2$  will imply a higher tendency to forming reciprocated ties.

**Degree Distribution** The following three effects are related to modeling the dynamics of the degree distribution. Since the data are directed graphs, three distinct aspects of the degree distribution are the variability of in-degrees, variability of out-degrees, and association between in- and out-degrees. Sometimes the term of preferential attachment is used [1,40] for the increased attractiveness of ties to nodes that already have high degrees. In our model these three aspects of preferential attachment can be expressed by including in the objective function terms depending on in- and out-degrees. The precise functional form for the effects is determined also by the requirement that the resulting model be amenable to statistical inference. Experience shows that using the square root of the degrees often leads to more stable estimation of the parameters than using the raw (untransformed) degrees. This suggests that, for many empirical networks, the models with squared roots of the degrees are better descriptions of reality than the models with untransformed degrees. Therefore only the models using the square roots are presented here.

### Popularity effect (square root measure)

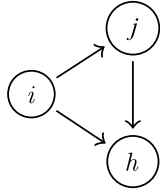
$s_{3i}(x) = \sum_j x_{ij} \sqrt{x_{+j}}$  This effect is defined by the sum of the square roots of indegrees of the others to whom  $i$  is tied. In other words, popularity of other actors is measured by the square root of their indegree. The root-popularity effect models the tendency to form ties to those actors who have high indegrees already (the Matthew effect in networks; see the chapter on [Social Network Analysis, Graph Theoretical Approaches](#) to). This will be reflected by the dispersion of the indegrees.

### Activity effect (square root measure)

$s_{4i}(x) = \sum_j x_{ij} \sqrt{x_{j+}}$  This effect is defined by the sum of the square roots of outdegrees of the others to whom  $i$  is tied. The effect models the tendency to form ties to those actors who have high outdegrees already. This will be reflected by the association between indegrees and outdegrees.

### Own outdegree, power 1.5, effect

$s_{4i}(x) = \sum_j x_{ij} \sqrt{x_{j+}} = x_{i+}^{1.5}$  The first expression show that each new tie has a ‘value’ equal to the actor’s outdegree, comparing to a unit ‘value’ for the outdegree effect. This effect expresses that actors who already have many outgoing ties have a higher propensity to establish new ties. This will lead to a greater dispersion of outdegrees.



Network Analysis, Longitudinal Methods of, Figure 1  
Transitive triplet

**Network Closure** The following two effects are related to network closure, also called transitivity or clustering: for friendship networks, this expresses that friends of friends tend to become friends. In addition to the two effects mentioned here, tendencies toward structural balance [7] and tendencies to avoid geodesic distances equal to 2 can also be implemented as effects that will lead to transitivity; also see [46,47].

**Transitive triplets effect**  $s_{5i}(x) = \sum_{j,h} x_{ij} x_{jh} x_{ih}$  This formula represents the number of transitive patterns in  $i$ 's ties as indicated in the figure below. A transitive triplet for actor  $i$  is a configuration  $(i, j, h)$  in which all three of the ties  $i \rightarrow j$ ,  $j \rightarrow h$ ,  $i \rightarrow h$  are present (and irrespective of whether there are also other ties between these three actors).

This models the tendency toward network closure, where (for a positive parameter) formation of the tie  $i \rightarrow h$  becomes increasingly likely when there are more indirect connections ('two-paths')  $i \rightarrow j \rightarrow h$ .

**Transitive ties effect**  $s_{6i}(x) = \sum_h x_{ih} \max_j (x_{ij} x_{jh})$

This is the number of actors  $j$  to whom  $i$  is directly as well as indirectly tied, i.e., for which there exists at least one  $h$  such that  $(i, j, h)$  is a transitive triplet.

The transitive triplets and transitive ties effects are two distinct ways of modeling tendencies toward network closure. For the effect on the probability of forming the tie  $i \rightarrow h$ , the number of two-paths  $i \rightarrow j \rightarrow h$  makes no difference for the transitive ties effect, as long as there is at least one indirect connection; for the transitive triplets effect the contribution to the objective function increases linearly with the number of two-paths.

**Three-cycle effect**  $s_{7i}(x) = \sum_{j,h} x_{ij} x_{jh} x_{hi}$  This is the number of three-cycles  $i \rightarrow j \rightarrow h \rightarrow i$  in which actor  $i$  is involved. This effect models the tendency toward forming three-cycles, which is the simplest form of generalized exchange [3,27] ( $i$  gives to  $j$ ,  $j$  gives to  $h$ , and  $h$  gives to  $i$ ) and which is opposed to hierarchy.

**Covariate Effects** Actor covariates  $v$  can influence the propensities to form or terminate ties in different ways,

because this propensity might be influenced by the value for the sender or the receiver of the tie, or by some combination of these two values.

**$v$ -related popularity**  $s_{8i}(x, v) = \sum_j x_{ij} v_j$  The  $v$ -related popularity effect is defined by the sum of the covariate over all actors to whom  $i$  is tied. Positive parameter values will imply that ties to actors with high  $v$  values are more attractive. This will lead to a tendency toward a correlation between  $v_i$  and the indegree of  $i$ .

**$v$ -related activity**  $s_{9i}(x, v) = v_i x_{i+}$  In words, this is  $v_i$  times  $i$ 's outdegree. Positive parameter values will imply that actors with high  $v$  values tend to make more ties. This will lead to a tendency toward a correlation between  $v_i$  and the outdegree of  $i$ .

**$v$ -related similarity**  $s_{10,i}(x, v) = \sum_j x_{ij} \text{sim}_v(i, j)$  Here  $\text{sim}_v(i, j)$  indicates the similarity between actors  $i$  and  $j$  defined by  $\text{sim}_v(i, j) = 1 - (|v_i - v_j|/\Delta)$ , where  $\Delta = \max_{h,k} |v_h - v_k|$  is the observed range of the covariate  $v$ . Thus, the effect is the sum of similarities between  $i$  and the others to whom he is tied. Positive parameter values will imply that there is a preference for ties between actors with similar values of  $v_i$  and  $v_j$ .

**$v$  ego-alter interaction**  $s_{11,i}(x, v) = \sum_j x_{ij} v_i v_j$  This product interaction is an alternative to the similarity effect for expressing that the propensity for a tie to exist depends on the combined values of the  $v$  for the sending actor ('ego',  $i$ ) and the receiving actor ('alter',  $j$ ).

**Main effect of**  $ws_{12,i}(x, w) = \sum_j x_{ij} w_{ij}$  For a dyadic covariate  $w$ , this is defined by the sum of the values of  $w_{ij}$  for all other actors  $j$  to whom  $i$  is tied.

It is clear that the list can be extended indefinitely, and that researchers have to make a limited choice reflecting theoretical and content-matter knowledge and interest. The potential complexity of network dynamics justifies to have many candidate effects that may be used to model the network dynamics. For instance, when a researcher wishes to estimate a model for testing whether there is a preference for choosing network partners with similar  $v$  values, while controlling for the tendency to have ties, and the tendencies for reciprocation and transitive closure, then effects  $s_1$ ,  $s_2$ ,  $s_5$  and/or  $s_6$ , and  $s_{10}$  should be included in the model (6). Since transitive closure could be expressed by effects  $s_5$  as well as  $s_6$ , there may be no strong prior arguments for choosing between these two 'control' effects, and empirical grounds could be used to choose either one or both. If there are grounds to suspect that the  $v_i$  values may also be associated with in- or outdegrees, the popularity and activity effects  $s_8$  and  $s_9$  may also be included.

### Specification of the Change Opportunity Process

For the model specification it should be noted that the ‘social time’ which determines the speed of change of the network is not necessarily the same as the physical time elapsing between consecutive observation moments. Given the absence of an extraneous definition of this ‘social time’, it is not a restriction to set to 1 the total time elapsed between each pair of consecutive observations. If there are  $M \geq 3$  observation moments, it is advisable to specify distinct rate parameters  $\rho_m$  governing the frequency of opportunities for change between  $t_m$  and  $t_{m+1}$ , and allow  $\rho_1, \rho_2$ , etc., to be different. If the change rate further is constant (independent of actors),  $\rho_m$  then represents the expected number of opportunities for change between  $t_m$  and  $t_{m+1}$ . This is the expected number per ordered pair  $(i, j)$  in the case of tie-based opportunities, and per actor  $i$  in the case of actor-based opportunities. The symbol  $\rho$  will denote the vector  $(\rho_1, \dots, \rho_{M-1})$ .

In the more general case the rate function can be defined, e.g., for the actor-based model, by a function depending on actor covariates and positional characteristics of the actors. When, for example, a dependence on one covariate  $v_i$  and the current out-degree  $x_{i+}^0$  is considered, a logarithmic link function could be used giving a model such as

$$\lambda_i(x^0, v) = \rho_m \exp(\alpha_1 v_i + \alpha_2 x_{i+}^0).$$

In general the symbol  $\alpha$  will be reserved for parameters indicating dependence of the rate function on covariates and network characteristics.

### Statistical Estimation and Testing

The most usual type of longitudinal network data is panel data, where for  $M \geq 2$  time points, an observation  $x(t_m)$  is available of the network on the same set  $\{1, \dots, n\}$  of actors.

These models can be simulated on computers in rather straightforward ways (the algorithm is written out in [47]). Parameter estimation, however, is more complicated, because the likelihood function or explicit probabilities can be computed only for uninteresting models. This section presents the Method of Moments (MoM) estimates proposed in [46]. Maximum Likelihood (ML) estimators are presented in [48]. In the current implementation, ML estimators are more time-consuming than MoM estimators, and can be used only for relatively small data sets. For the more straightforward models (dynamics of networks only, no endowment functions), the MoM method is hardly less efficient than the ML method. For more complicated mod-

els, where it is important to squeeze every bit of information out of the data, it can be useful to employ ML methods. In the following description of the estimation method, the parameter vector  $(\rho, \alpha, \beta)$  is denoted by  $\theta$ .

It is undesirable to make the restrictive assumption that the distribution of the process is stationary. Instead, for each observation moment  $t_m$  ( $m = 1, \dots, M-1$ ) the observed network  $\mathbf{x}(t_m)$  can be used as a conditioning event for the distribution of  $\mathbf{X}(t_{m+1})$ . The Method of Moments requires that a vector of statistics  $U_{m+1} = U(\mathbf{X}(t_m), \mathbf{X}(t_{m+1}))$  is utilized, such that the expected value

$$E_\theta \{U(\mathbf{X}(t_m), \mathbf{X}(t_{m+1})) \mid \mathbf{X}(t_m) = \mathbf{x}(t_m)\}$$

is sensitive to the parameter  $\theta$ . Given the conditioning on the preceding observation, the moment equations, or estimating equations, can then be written as

$$\sum_{m=1}^{M-1} E_\theta \{U(\mathbf{X}(t_m), \mathbf{X}(t_{m+1})) \mid \mathbf{X}(t_m) = \mathbf{x}(t_m)\} = \sum_{m=1}^{M-1} U(\mathbf{x}(t_m), \mathbf{x}(t_{m+1})). \quad (7)$$

It turns out that suitable statistics are the following. The number of changed ties between consecutive observations,

$$\sum_{i,j} |X_{ij}(t_{m+1}) - X_{ij}(t_m)|,$$

is especially sensitive to the rate of change  $\rho_m$ . A vector of statistics sensitive especially to  $\beta$  is the sum of the individual objective functions

$$\sum_i f_i(\mathbf{X}(t_{m+1})).$$

To solve the estimating equation (7), in the absence of ways to calculate analytically the expected values, stochastic approximation methods can be used. Variants of the Robbins–Monro [9,42] algorithm have been used with good success. This is a stochastic iteration method which produces a sequence of estimates  $\theta^{(N)}$  which is intended to converge to the solution of (7), and which works here as follows. For a given provisional estimate  $\theta^{(N)}$ , the model is simulated so that for each  $m = 1, \dots, M-1$ , a simulated random draw is obtained from the conditional distribution of  $\mathbf{X}(t_{m+1})$  conditional on  $\mathbf{X}(t_m) = \mathbf{x}(t_m)$ . This simulated network is denoted  $\mathbf{X}^{(N)}(t_{m+1})$ . Denote  $U_m^{(N)} = U(\mathbf{x}(t_m), \mathbf{X}^{(N)}(t_{m+1}))$  and  $U^{(N)} = \sum_{m=1}^{M-1} U_m^{(N)}$ ,

and let  $u^{\text{obs}}$  be the right-hand side of (7). Then the iteration step in the Robbins–Monro algorithm for obtaining the Method of Moments estimate is given by

$$\theta^{(N+1)} = \theta^{(N)} - a_N D^{-1} (U^{(N)} - u^{\text{obs}}), \quad (8)$$

where  $D$  is a suitable matrix and  $a_N$  a sequence of positive constants tending to 0. Tuning details of the algorithm, including the choices of  $D$  and  $a_N$ , are given in [46]. The experience with the convergence of this algorithm is quite good. The standard errors can be computed using the standard formulae of standard errors for the Method of Moments, based on the delta method, and applying simulation methods; such simulation methods are discussed in [44]. Bayesian estimators for these models are presented in [24] and Maximum Likelihood estimators in [48].

### Example: Dynamics of Adolescent Friendship

As an example, the adolescent friendship network is considered of a year cohort at a secondary school in Glasgow (Scotland), studied in the Teenage Friends and Lifestyle Study [36,38]. This data set was collected at three measurement points  $t_1, t_2, t_3$  in 1995–1997, at intervals of roughly one year, starting when the pupils were 12–13 years old. Here the network is studied that is formed by the 129 (out of 160) pupils who were present at all three measurement waves. Sex (boys scored as 1, girls as 2) and drinking behavior are used as actor variables, drinking (alcohol consumption) being measured on a 5-point scale ranging from 1 (not at all) to 5 (more than one a week). Both variables are centered around the mean, which is 1.43 for sex and 2.60 for drinking (averaged over  $t_1$  and  $t_2$ ). The data set is used as an illustration; more wide-ranging analyses are presented in [37,52,53].

Three models are presented. Rate parameters are assumed constant within periods between observation moments, and the duration of the periods is (arbitrarily but without loss of generality) set at 1. The first model contains only the most basic dyadic and triadic effects: outdegree, reciprocity, transitive triplets, and three-cycles. The second adds to this the three effects to model more precisely the degree distribution. The third model adds to these structural effects the effects of two covariates: gender and alcohol consumption.

Parameter estimates are approximately unbiased and normally distributed; for this assertion there is no mathematical proof yet, but it is supported by simulation studies. Therefore, effects can be tested by referring the studentized estimates (or  $t$ -ratios, i. e., estimate divided by standard error) to a standard normal distribution. When the  $t$ -ratio

exceeds 2 in absolute value, the effect can be interpreted as being significant at the significance level of 5 %.

The table present first the estimated rate parameters. These indicate that the pupils had about 12 opportunities for changes in the first period ( $t_1 - t_2$ ), and about 9 in the second period ( $t_2 - t_3$ ).

The parameters for the objective function, which the table shows next, are more important for the interpretation. There are strong tendencies toward reciprocity and transitivity, and a tendency away from three-cycles. This is the case in all three models, although the parameter estimates are slightly different. This indicates that the structural features of transitive closure and of local hierarchy (the interpretation of the negative three-cycle effect) cannot be ‘explained away’ by tendencies in tie formation and dissolution that are associated to degrees, sex, or alcohol consumption.

The degree effects in Models 2 as well as 3 show that there is a positive popularity effect – with a borderline significance at the 5% level in Model 3; there is a negative activity effect; and a negative effect of own outdegree raised to the power 1.5. The positive popularity effect suggests that differential values in in-degrees tend to be self-sustaining, leading to rather strongly dispersed in-degrees. The other two, negative, degree-related effects indicate that differences in out-degrees, as well as correlations between in- and out-degrees, tend to be self-correcting, leading to relatively low dispersions of out-degrees and low in-degree – out-degree correlations.

Of the three gender-related effects, only the similarity effect is significant. In this age range, a strong preference for same-sex friendships is to be expected (the interest which perhaps exists in the other sex is not reported as friendship). For alcohol consumption, the interaction effect shows that those who drink more themselves have a higher preference for friends who also drink more; and the activity effect shows that those drinking more tend to mention less friends, for friends of average drinking habits (where the contribution of the interaction is nil). This is most clearly expressed by jointly considering the three contributions related to drinking behavior that can be made by the tie  $i \rightarrow j$ . In a formula, this is represented by the coefficient of the variable  $y_{ij}$  in the objective function. Recall that the actor variables are centered around the mean. Using the formulae for effects  $s_8, s_9$ , and  $s_{11}$  and filling in the coefficients in Table 1 yields

$$-0.026 (v_j - \bar{v}) - 0.086 (v_i - \bar{v}) + 0.107 (v_j - \bar{v})(v_i - \bar{v}),$$

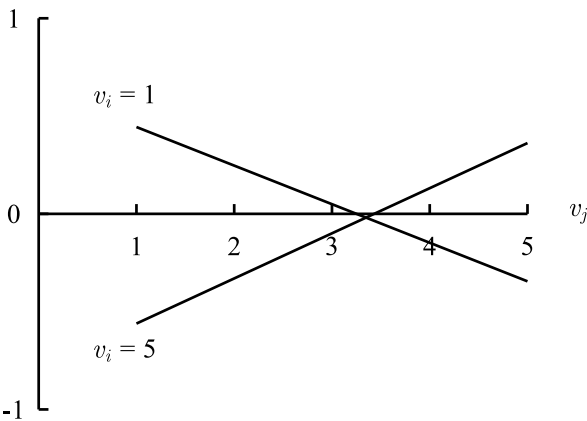
with  $\bar{v} = 2.60$ . The following picture illustrates the contributions to the objective function made by ego’s drinking behavior  $v_i$  and alter’s drinking behavior  $v_j$ .



**Network Analysis, Longitudinal Methods of, Table 1**  
**Parameter estimates for modeling evolution of friendship network, Glasgow school cohort. Standard errors between parentheses**

Effect	Model 1		Model 2		Model 3	
	par.	(s.e.)	par.	(s.e.)	par.	(s.e.)
Rate 1	11.82	(1.00)	12.05	(1.15)	12.19	(1.13)
Rate 2	9.23	(0.83)	9.05	(0.79)	9.09	(0.77)
Outdegree	-2.697	(0.047)	-0.16	(0.34)	-0.18	(0.41)
Reciprocity	2.38	(0.10)	2.48	(0.12)	2.22	(0.12)
Transitive triplets	0.459	(0.033)	0.569	(0.036)	0.544	(0.034)
Three -cycles	-0.57	(0.10)	-0.55	(0.11)	-0.44	(0.11)
Popularity (sq. root)	-		0.223	(0.095)	0.184	(0.093)
Activity (sq. root)	-		-0.89	(0.14)	-0.92	(0.17)
Own outdegree, power 1.5	-		-0.560	(0.084)	-0.587	(0.094)
Sex (F) popularity	-		-		-0.16	(0.11)
Sex (F) activity	-		-		0.12	(0.12)
Sex similarity	-		-		0.904	(0.097)
Drinking popularity	-		-		-0.026	(0.030)
Drinking activity	-		-		-0.086	(0.038)
Drinking ego × alter	-		-		0.107	(0.024)

It can be concluded that those who do not drink alcohol ( $v_i = 1$ ), prefer friends who drink no or little alcohol, while the reverse is true for those who drink a lot of alcohol. The range of this contribution is 0.8 for those with  $v_i = 1$  and 0.9 for those with  $v_i = 5$ , which is comparable to the value 0.9 of the gender similarity effect. Thus, for those with the highest and the lowest values of alcohol use, the importance of the alcohol use of potential friends, when comparing potential friends with the minimum and the maximum alcohol use, is approximately as great as the importance of their gender. On the other hand, for those with medium values of alcohol use, the alcohol use of potential friends plays virtually no role at all.



**Network Analysis, Longitudinal Methods of, Figure 2**  
**Contributions of drinking behaviors  $v_i, v_j$  to the objective function for friendship**

### Models for the Co-evolution of Networks and Behavior

The importance of networks derives, for an important part, from the effects of networks on the behavior and performance of the actors. This can be, for example, because of influence between friends or between collaborating partners [16], because of exchange of resources [28], or because of structural advantages [4].

A few examples drawn from the many studies in this field are concerned with influence of friends of adolescents on smoking behavior [14,37,52], effects of acquaintances on labor market outcomes [17], competitive advantage of firms [18], effects of friends on delinquency [6,19], effects of position in patent citation networks on growth rates of companies [39], and job performance [51].

This type of changing individual attributes, which might range from, e.g., behavioral tendencies and attitudes of individuals to performance of companies, will be briefly referred to as ‘behavior’. When the ties in the network are influenced by the behavior and the behavior, in turn, is influenced by the network, a mutual feedback arises between the network and the behavior. The network structure of the ties between the actors together with their behavior constitute the endogenously changing environment for each of the actors [61]. Thus, this approach is well-suited to study macro-micro-macro questions of the type discussed by [11].

The vector of attributes for actor  $i$  at time  $t$  is denoted

$$z_i(t) = (z_{i1}(t), \dots, z_{iH}(t))$$

where  $z_{ih}(t)$  denotes the  $h$ th attribute of actor  $i$ . For the  $n$  actors these are stacked in the matrix  $z(t)$ , which is regarded as the outcome of a random matrix  $Z(t)$ .

To model the co-evolution of networks and behavior, where the network and the behavior influence one another dynamically, the stochastic process  $(X(t), Z(t))$  is considered. This is treated in quite the same way as the stochastic process  $X(t)$  was treated above. It is assumed that  $z_i(t)$  represents an enduring, but changeable, state of actor  $i$  rather than a momentary behavior, so that it is a sensible approximation to assume that  $(X(t), Z(t))$  is a Markov process in continuous time. Now the change probabilities of  $X(t)$  will depend on the current state of  $X(t)$  as well as  $Z(t)$ . The fact that network change is co-determined by behavior is called *behavior-dependent selection*, and will often be referred to briefly as *selection*. Similarly, the change probabilities of  $Z(t)$  will depend on the current state of  $Z(t)$  as well as  $X(t)$ , and this will be called *influence*. This terminology was also used, e. g., by [14].

To remain close to the framework for network modeling, it is assumed that all of the behavior variables  $z_{ih}$  are measured on a discrete ordinal scale, with values coded as consecutive integers  $\{0, 1, \dots, M_h\}$  for some  $M_h \geq 1$ . The analogue of the simplifying assumption (3) made above for network change is the following. This decomposes the change between consecutive observations  $(x(t_m), z(t_m))$  and  $(x(t_{m+1}), z(t_{m+1}))$  into a sequence of the smallest possible steps.

3'. At any given moment  $t$ , no more than one of all the variables  $X_{ij}(t), Z_{ih}(t)$  can change. When  $Z_{ih}(t)$  changes, at any given instant it can change only to an immediate neighboring value, i. e., by a decrease or increase of 1 (permitted only if this step does not take  $Z_{ih}$  outside of the permitted range from 0 to  $M_h$ ).

This means that there is no direct coordination between changes in ties and changes in behavior, and the dependence between networks and behavior is brought about because both react to each other.

The model for the network is just like it was above, the rate function and objective function now being denoted by  $\lambda^X$  and  $f_i^X$ , and allowed to depend on the current behavior  $Z(t)$ , to represent behavior-dependent selection. For each of the dependent behavior variables  $Z_h$  there also is a rate function  $\lambda_i^{Z_h}$  driving the frequency of changes and an objective function  $f_i^{Z_h}$  defining the probabilities of behavior changes when there is an opportunity of change. Their dependence on the current network will represent influence.

For each actor  $i$ , opportunities to change behavior  $Z_{ih}$  occur according to a Poisson process with rate  $\lambda_i^{Z_h}$ .

When actor  $i$  has the opportunity to change behavior  $Z_{ih}$  given that currently  $(X(t), Z(t)) = (x^0, z^0)$ , there are three possible new states  $(x^0, z)$ , where either  $z = z^0$  or the only difference between  $z$  and  $z^0$  is that  $z_{ih} = z_{ih}^0 - 1$  or  $z_{ih} = z_{ih}^0 + 1$ ; unless one of these values is outside the range  $\{0, \dots, M_h\}$ , in which case there are only the two remaining possible new states.

The probabilities of going from state  $(x^0, z^0)$  to state  $(x^0, z)$  are proportional to  $\exp(f_i^{Z_h}(z^0, z, x^0, v, w))$ .

The change probabilities in this model are given by

$$\begin{aligned} &P\{Z(t) \text{ changes to } z \mid i \text{ has an opportunity to change } \\ &Z_{ih} \text{ at time } t, X(t) = x^0, Z(t) = z^0\} \\ &= \frac{\exp(f_i^{Z_h}(z^0, z, x^0, v, w))}{\sum_{z' \in C^{Z_h}(z^0)} \exp(f_i^{Z_h}(z^0, z', x^0, v, w))}, \quad (9) \end{aligned}$$

where  $C^{Z_h}(z^0)$  is the set of two or three permitted new values for  $Z_h$ .

The effects in the objective function  $f_i^X$  for network changes now are denoted by  $s_{ki}^X$  with weights  $\beta_k^X$ , and likewise the objective function for behavior changes is assumed to be expressed as a linear combination of effects weighted by parameters,

$$f_i^{Z_h}(z^0, z, x^0, v, w) = \sum_k \beta_k^{Z_h} s_{ki}^{Z_h}(z^0, z, x^0, v, w). \quad (10)$$

A set of possible effects that could be included in the objective function for behavior are the following. For simplicity of notation, it is assumed that there is only one behavior variable, so that the index  $h$  can be dropped from the notation. The first two effects are used to define the shape of the objective function as a function of  $z_i$ , and other terms depending on  $z_i$  could be added.

**Linear Shape Effect**  $s_{i1}^Z(z) = z_i$

**Squared Shape Effect**  $s_{i2}^Z(z) = z_i^2$  The linear and squared shape effects together define what could be regarded as a quadratic preference function on the behavior,  $\beta_1^Z z_i + \beta_2^Z z_i^2$ . The word 'preference function' is used with some reluctance, because it is used here only as an easy shorthand term for the combined short-term result of preferences and constraints, depending only on the actor's behavior  $z_i$  itself, net of the other terms in

the behavior objective function. If  $\beta_2 < 0$  this is a uni-modal function of  $z_i$ . If  $\beta_2 > 0$ , on the other hand, and if the minimum of the quadratic function is assumed within the range of the behavior variable, then the behavior is drawn to the extremes of the range, with actors already low on  $z_i$  being drawn to low values and actors already high on  $z_i$  being drawn to high values. This can represent, e.g., addictive behavior.

Other nonlinear functions of  $z_i$  could, of course, also be included.

**Indegree effect**  $s_{i3}^Z(z, x) = z_i x_{i+}$  This represents that actors with a high indegree ('popular' actors) have a higher tendency toward high values of the behavior.

**Outdegree effect**  $s_{i4}^Z(z, x) = z_i x_{i+}$  This represents that actors with a high outdegree ('active' actors) have a higher tendency toward high values of the behavior.

**Total similarity effect**  $s_{i5}^Z(z, x) = \sum_j x_{ij} \text{sim}_z(i, j)$  The dyadic similarity  $\text{sim}_z$  is as defined above, now applied to the dependent behavior  $Z$ . The total similarity effects adds the similarity values between  $i$  and the actors toward whom  $i$  has a tie. This is the primary representation of social influence: the preference for behavior which is close to that of one's network members.

**Average alter effect**  $s_{i6}^Z(z, x) = z_i x_{i+}^{-1} \sum_j x_{ij} z_j$  The coefficient of  $z_i$  is the average behavior of the actors to whom  $i$  is tied ( $i$ 's 'alters',  $j$ ). The coefficient is defined as 0 when the outdegree  $x_{i+}$  is 0. This is another representation of social influence: the preference for behavior depends on the average behavior of one's network members.

The parameter estimation for this model is discussed in [49].

### Example: Co-evolution of Adolescent Friendship and Alcohol Use

As an example for the co-evolution of networks and behavior the data of the Glasgow school cohort is used again, but now the alcohol consumption is used as a dependent, or endogenous, variable. In the treatment in Table 1 the alcohol consumption was used as an exogenous variable, i.e., its values were accepted as if determined by processes independent of the network. Now we follow a co-evolution approach where it is assumed that the dynamics in alcohol consumption can be co-determined by the network just as the network dynamics can be co-determined by the alcohol consumption.

To obtain interpretations of the numerical parameter values, it must be noted that this table refers to the centered actor variables. The raw variable ranges from 1 to 5, with averages at the three observations rising from 2.5

### Network Analysis, Longitudinal Methods of, Table 2

Parameter estimates for modeling co-evolution of friendship network and alcohol consumption, Glasgow school cohort

Effect	par.	(s.e.)
<i>Friendship dynamics</i>		
Rate 1	12.13	(1.26)
Rate 2	9.09	(0.97)
Outdegree	−0.22	(0.43)
Reciprocity	2.19	(0.12)
Transitive triplets	0.529	(0.036)
Three-cycles	−0.42	(0.10)
Popularity (sq. root)	0.193	(0.086)
Activity (sq. root)	−0.92	(0.15)
Own outdegree, power 1.5	−0.581	(0.099)
Sex (F) popularity	−0.16	(0.10)
Sex (F) activity	0.12	(0.13)
Sex similarity	0.90	(0.11)
Drinking popularity	−0.026	(0.043)
Drinking activity	−0.118	(0.057)
Drinking ego × alter	0.166	(0.043)
<i>Alcohol consumption dynamics</i>		
Rate 1	1.48	(0.25)
Rate 2	2.22	(0.37)
Linear shape	0.36	(0.42)
Squared shape	−0.34	(0.14)
Indegree	0.07	(0.12)
Outdegree	−0.08	(0.17)
Sex (F)	−0.02	(0.23)
Average alter	0.83	(0.35)

through 2.7 to 3.1. The overall mean, 2.8, is subtracted from the observations. We shall denote by  $z_i$  the raw value of actor  $i$ 's alcohol consumption scored 1–5, by  $\bar{z}$  the average equal to 2.8, and by  $\bar{z}_i$  the average of the alcohol consumption of  $i$ 's friends.

The interpretation of the friendship dynamics is quite the same as in the analysis where alcohol was treated as an exogenous variable. For the dynamics of alcohol consumption, the indegree, outdegree, and sex of the actor do not seem to have an important influence. The effect of the average drinking behavior of the friends of the focal actor does have a significant effect, however, with a  $t$ -value of  $0.83/0.35 = 2.4$ . When we ignore the small and non-significant effects of outdegree, indegree, and sex, the remaining part of the objective function for drinking behavior is

$$0.36(z_i - \bar{z}) + 0.83(\bar{z}_i - \bar{z})(z_i - \bar{z}) - 0.34(z_i - \bar{z})^2 \\ = (0.36 + 0.83(\bar{z}_i - \bar{z}))(z_i - \bar{z}) - 0.34(z_i - \bar{z})^2.$$

This is a quadratic function, unimodal, with a maximum at

$$z_i = \bar{z} + \frac{0.83}{2 \times 0.34} (\bar{z}_i - \bar{z}) = -0.62 + 1.22 \bar{z}_i .$$

Taking account of the integer values of  $z_i$ , this implies that those with friends with the smallest possible average of smoking behavior  $\bar{z}_i = 1$  are drawn towards the ‘preferred’ value of 1 themselves, while those having friends who have the highest possible average  $\bar{z}_i = 5$  are themselves also drawn toward this value 5 as a ‘preferred’ value. It can be concluded that the data provide support for the existence, in the co-evolution of friendship and drinking tendencies, of selection (tendency to choose friends with similar behavior) as well as influence (tendency to change behavior in the direction of friends’ behavior).

## Extensions

The basic model specifications defined above can be extended in various ways. Above we already mentioned the possibility to let the rates of change depend on covariates or on current network structure. Another possibility is to introduce an asymmetry between the values of ties when they are formed and their values when they are lost. E.g., for friendship dynamics, there is theoretical and empirical evidence that the additional ‘value’ of a tie added by its being reciprocated is higher when considering a potential loss of the tie than when considering the potential new formation of the tie. This asymmetry can be modeled by endowment functions, see [49]. Technically, this means that the effects  $s_{ki}(x^0, x, v, w)$  used in (6) depend not only on the new state  $x$  but, unlike the examples given above, also on the preceding state  $x^0$ . For example, the endowment effect of a tie being reciprocal is expressed by

$$s_{ki} = \sum_j x_{ij}^0 x_{ij} x_{ji} ,$$

which is sensitive to the number of  $i$ ’s reciprocal ties only when they are candidates to being terminated and not when they are candidates to being created.

Similarly, going upward on a behavior variable might be not the opposite of going downward, which can be modeled by endowment effects in the objective function for behavior.

## Future Directions

Although the statistical modeling of network dynamics started already with [22] and [59], this area has been in rapid development only since recent years. Much work remains to be done, however, to extend these methods to

other types of network data and to study their properties. The lists of effects that can be included in the objective functions illustrate the flexibility of this model and its adaptability to research questions and network data. The availability of methods for analysis of network panel data has been a stimulus also for the further collection of such data.

Plausible models and good methods for parameter estimation and testing have now been developed, as summarized in this chapter, and they are available in the SIENA (*Simulation Investigation for Empirical Network Analysis*) program. This program is available as freeware with additional material on the website <http://stat.gamma.rug.nl/snijders/siena.html> and has an extensive manual [50]. The examples presented here were analyzed using this program.

The tests used in the examples, based on studentized parameter estimates, can be regarded as Wald-type tests. Some limited simulation studies have supported the validity of these tests. Score-type tests associated to the Method of Moments estimators were developed in [43]. The examples in this paper underline the need for methods to assess fit of models, and also to compare non-nested models. This could be done formally based on estimated likelihoods, or informally based on the comparison of observed and expected values of relevant statistics that are not used for parameter estimation.

The open question of assessing fit also invites speculation about the robustness of the results against the use of models of which the fit is not beyond doubt. There is an inherent tension between the complexity of processes of network dynamics, and the limited amount of data that can in practice be observed concerning these processes. One issue is that the models proposed here are Markov processes. For two-wave data sets there are no clear alternatives to making such an assumption, but the assumption is certainly debatable. Including more information in the state space (by using covariates, by considering valued rather than dichotomous ties, etc.) may relax the doubts concerning such an assumption.

Another issue is the difference between the tie-oriented and actor-oriented models. Which type of model is to be preferred is a matter both of social science theory and of empirical fit. It will be important to know, supported by simulation studies and/or mathematical results, the extent to which results based on particular models for network dynamics are robust to deviations from the precise assumptions made. In addition, it will be useful to develop still other models, e.g., models accounting for actor heterogeneity (like were developed for non-longitudinal network data, e.g., by [20,35,58]) or measurement error.

Other open questions are about mathematical properties of the estimators and tests proposed. Simulation studies support the conjecture that the Method of Moments estimators have asymptotically normal distributions, but this has not been proven. It is unknown if the solution to the moment equation (7), under certain conditions, is unique. Similar questions can be asked about the Maximum Likelihood estimators. All this indicates that there is ample scope for future work on methods of statistical inference for network dynamics.

## Bibliography

### Primary Literature

- Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
- Bartholomew DJ (1982) *Stochastic Models for Social Processes*, 3rd edn. Wiley, New York
- Bearman PS (1997) Generalized exchange. *Am J Sociol* 102:1383–1415
- Burt RS (1992) *Structural Holes*. Harvard University Press, Cambridge
- Brass DJ, Galaskiewicz J, Greve HR, Tsai W (2004) Taking stock of networks and organizations: a multilevel perspective. *Acad Manag J* 47:795–817
- Burk WJ, Steglich CEG, Snijders TAB (2007) Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. *Int J Behav Dev* 31:397–404
- Cartwright D, Harary F (1956) *Structural Balance: A Generalization of Heiders Theory*. *Psychol Rev* 63:277–292
- Checkley M, Steglich CEG (2007) Partners in Power: Job Mobility and Dynamic Deal-Making. *Eur Manag Rev* 4:161–171
- Chen H-F (2002) *Stochastic Approximation and its Applications*. Kluwer, Dordrecht
- Coleman JS (1964) *Introduction to Mathematical Sociology*. The Free Press of Glencoe, New York
- Coleman JS (1990) *Foundations of Social Theory*. Belknap Press of Harvard University Press, Cambridge/London
- de Federico de la Rua A (2003) La dinamica de las redes de amistad. La eleccion de amigos en el programa Erasmus. REDES 4.3, <http://revista-redes.rediris.es>. Accessed 10 Aug 2007
- de Nooy W (2002) The dynamics of artistic prestige. *Poetics* 30:147–167
- Ennett ST, Bauman KE (1994) The Contribution of Influence and Selection to Adolescent Peer Group Homogeneity: The Case of Adolescent Cigarette Smoking. *J Personal Soc Psychol* 67:653–63
- Doreian P, Stokman FN (eds) (1997) *Evolution of Social Networks*. Gordon and Breach, Amsterdam
- Friedkin NE (1998) *A Structural Theory of Social Influence*. Cambridge University Press, Cambridge
- Granovetter MS (1973) The strength of weak ties. *Am J Sociol* 78:1360–1380
- Gulati R, Nohria N, Zaheer A (2000) Strategic Networks. *Strateg Manag J* 21:203–215
- Haynie DL (2001) Delinquent Peers Revisited: Does Network Structure Matter? *Am J Sociol* 106:1013–1057
- Hoff PD, Raftery AE, Handcock MS (2002) Latent Space Approaches to Social Network Analysis. *J Am Stat Assoc* 97:1090–1098
- Holland PW, Leinhardt S (1975) Local structure in social networks. *Sociol Methodol* – 1976 1–45
- Holland PW, Leinhardt S (1977) A dynamic model for social networks. *J Math Sociol* 5:5–20
- Hosmer D, Lemeshow S (2000) *Applied logistic regression*, 2nd edn. Wiley-Interscience, New York
- Koskinen JH, Snijders TAB (2007) Bayesian inference for dynamic social network data. *J Stat Plan Inference* 137:3930–3938
- Kossinets G, Watts DJ (2006) Empirical Analysis of an Evolving Social Network. *Science* 311:88–90
- Leenders RTAJ (1995) Models for network dynamics: A Markovian framework. *J Math Sociol* 20:1–21
- Lévi-Strauss Ce (1969; orig. 1947) *The Elementary Structures of Kinship*. Beacon Press, Boston
- Lin N, Cook K, Burt RS (eds) (2001) *Social Capital. Theory and Research*. Aldine de Gruyter, New York
- Long JS (1997) *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, Thousand Oaks
- Maddala GS (1983) *Limited-dependent and Qualitative Variables in Econometrics*, 3rd edn. Cambridge University Press, Cambridge
- Newcomb TM (1962) Student Peer-Group Influence. In: Sanford N (ed) *The American College: A Psychological and Social Interpretation of the Higher Learning*. Wiley, New York
- Newman MEJ, Barkema GT (1999) Monte Carlo methods in Statistical Physics. Clarendon Press, Oxford
- Newman MEJ, Watts DJ, Strogatz SH (2002) Random graph models of social networks. *Proceedings of the National Academy of Sciences USA*, 99, 2566–2572, National Academy of Sciences, Washington
- Norris JR (1997) *Markov Chains*. Cambridge University Press, Cambridge
- Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc* 96:1077–1087
- Pearson MA, Michell L (2000) Smoke Rings: Social network analysis of friendship groups, smoking and drug-taking. *Drugs Educ Prev Policy* 7:21–37
- Pearson M, Steglich C, Snijders T (2006) Homophily and assimilation among sport-active adolescent substance users. *Connections* 27(1):47–63
- Pearson M, West P (2003) Drifting Smoke Rings: Social Network Analysis and Markov Processes in a Longitudinal Study of Friendship Groups and Risk-Taking. *Connections* 25(2): 59–76
- Podolny JM, Stuart TE, Hannan MT (1997) Networks, knowledge, and niches: competition in the worldwide semiconductor industry, 1984–1991. *Am J Sociol* 102:659–689
- de Solla Price D (1976) A general theory of bibliometric and other advantage processes. *J Am Soc Inf Sci* 27:292–306
- Robins GL, Woolcock J, Pattison P (2005) Small and other worlds: Global network structures from local processes. *Am J Sociol* 110:894–936
- Robbins H, Monro S (1951) A stochastic approximation method. *Ann Math Stat* 22:400–407
- Schweinberger M (2007) Statistical modeling of digraph panel data: Goodness-of-fit. Submitted for publication



44. Schweinberger M, Snijders TAB (2006) Markov models for di-graph panel data: Monte Carlo-based derivative estimation. *Comput Stat Data Anal* 51:4465–4483
45. Snijders TAB (1996) Stochastic actor-oriented dynamic network analysis. *J Math Sociol* 21:149–172
46. Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel M, Becker M (eds) *Sociological Methodology*. Basil Blackwell, Boston and London, pp 361–395
47. Snijders TAB (2005) Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S (eds) *Models and Methods in Social Network Analysis*. Cambridge University Press, New York
48. Snijders TAB, Koskinen JH, Schweinberger M (2007) Maximum Likelihood Estimation for Social Network Dynamics. Submitted for publication
49. Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) *Longitudinal models in the behavioral and related sciences*. Lawrence Erlbaum, Mahwah, pp 41–71
50. Snijders TAB, Steglich CEG, Schweinberger M, Huisman M (2007) Manual for SIENA version 3. ICS, University of Groningen; Groningen, Department of Statistics, University of Oxford, Oxford <http://stat.gamma.rug.nl/snijders/siena.html>. Accessed 4 Dec 2007
51. Sparrowe RT, Liden RC, Wayne SJ, Kraimer ML (2001) Social Networks and the performance of individuals and groups. *Acad Manag J* 44:316–325
52. Steglich CEG, Snijders TAB, Pearson M (2007) Dynamic networks and behavior: Separating selection from influence. Submitted for publication
53. Steglich CEG, Snijders TAB, West P (2006) Applying SIENA: An Illustrative Analysis of the Coevolution of Adolescents' Friendship Networks, Taste in Music, and Alcohol Consumption. *Methodology* 2:48–56
54. Udehn L (2002) The changing face of methodological individualism. *Ann Rev Sociol* 8:479–507
55. van de Bunt GG, van Duijn MAJ, Snijders TAB (1999) Friendship networks through time: An actor-oriented statistical network model. *Comput Math Organ Theory* 5:167–192
56. van de Bunt GG, Wittek RPM, de Klepper MC (2005) The Evolution of Intra-Organizational Trust Networks; The Case of a German Paper Factory: An Empirical Test of Six Trust Mechanisms. *Int Sociol* 20:339–369
57. van Duijn MAJ, Zeggelink EPH, Huisman M, Stokman FN, Wasseur FW (2003) Evolution of Sociology Freshmen into a Friendship Network. *J Math Sociol* 27:153–191
58. van Duijn MAJ, Snijders TAB, Zijlstra BH (2004)  $p_2$ : a random effects model with covariates for directed graphs. *Stat Neerlandica* 58:234–254
59. Wasserman S (1980) Analyzing social networks as stochastic processes. *J Am Stat Assoc* 75:280–294
60. Wippler R (1978) The structural-individualistic approach in Dutch sociology. *Neth J Sociol* 4:135–155
61. Zeggelink EPH (1994) Dynamics of structure: an individual oriented approach. *Soc Netw* 16:295–333

### Further Reading

For further reading, basic concepts of continuous-time Markov processes can be found in [34]. The basic definition of the model presented here and of the statistical estimation meth-

ods for network dynamics based on panel data can be studied in [46,47]. The approach to the co-evolution of networks and behavior is presented in [49,52]. Some examples of the methods presented in this chapter can be found in [6,53,55,56,57].

## Networks, Flexibility and Mobility in

MICHAEL F. THORPE

Arizona State University, Tempe, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Flexibility  
 Maxwell Counting  
 Enumeration Methods  
 Molecular Framework Conjecture  
 Geometrical Simulation  
 Future Directions  
 Bibliography

### Glossary

**Atypical graph** A special graph with symmetry, such as parallel lines – in contrast to a generic graph.

**Covalent bond** When a covalent bond exists between two atoms in a molecule, the bond length is fixed and independent of the environment. The bond angle is also usually fixed between two covalent bonds that share a common atom.

**Generic graph** A graph with arbitrary positions of the vertices and associated edges.

**Graph** A set of vertices connected by edges.

**Dihedral angle rotation** A rotation about a covalent bond connecting two atoms in a larger molecule.

**Geometrical simulation** A technique that allows the motion of the flexible parts of a network to be determined while obeying all the constraints, both equalities and inequalities.

**Hyperstatic** When the number of degrees of freedom is less than the number of constraints plus the number of rigid body motions.

**Hypostatic** When the number of degrees of freedom is greater than the number of constraints plus the number of rigid body motions.

**Isostatic** When the number of degrees of freedom is equal to the number of constraints plus the number of rigid body motions.

**Macromolecule** A large number of atoms connected by covalent bonds.

**Maxwell count** A global estimate of the number of degrees of freedom that assumes that every edge (constraint) is independent.

**Molecular dynamics** The solution of the classical equations of motion for a system of particles moving under a given potential.

**Network glass** Non crystalline network made up of atoms that form covalent bonds with each other.

**Nuclear magnetic resonance** An experimental technique that can be used to probe the different conformational states of a molecule.

**Pebble game** A procedure for book keeping between degrees of freedom and constraints that allows a rigid region decomposition to be performed on a graph.

**Protein** A large macromolecule made up of amino acids linked together to form a polypeptide chain, which folds into a compact structure that has a biological function.

**Rigid region decomposition** The division of a graph into rigid regions, both with and without redundant edges, and the flexible regions between them.

**Redundant edge** An edge that can be removed without changing the rigid regions in the graph.

## Definition of the Subject

A network can be represented by a graph that can be either rigid or flexible depending on the number and distribution of the constraints. A constraint is a length that is fixed. One example of a network is a collection of points in space connected by edges. If the number of edges is large enough and homogeneous enough in distribution, then the system is a rigid body, and no motion is possible, except for translations and rotations of the whole network. If a sufficient number of edges are removed, then parts of the systems can become flexible, which means that local motion is possible while maintaining all the constraints. Another manifestation of such a system is a collection of bodies in space connected by bars. Once it is determined that there are flexible parts in the graph, the actual motion, or mobility, can be found. Such graphs were first studied by Maxwell [1] in the 1860s using ideas previously developed by Lagrange [2] in the 1780s. Exact enumeration methods were developed later, starting with Laman [3] in 1970 for two-dimensional systems and can be extended in three-dimensional graphs in some cases [4,5]. The mobility of flexible graphs can be studied using geometrical simulation [6].

This general approach, involving flexibility and mobility, has important applications in complex graphs (networks) where it is difficult to solve the classical equations of motion [7] within a potential energy landscape [8], and insight can be gained by approximating the energy landscape by a set of appropriate constraints. These constraints can be either equalities as described above, which are important in determining the flexibility of the network, or equalities augmented by inequalities which determine the subsequent mobility. Two important applications at the atomic and molecular level are network glasses and proteins.

## Introduction

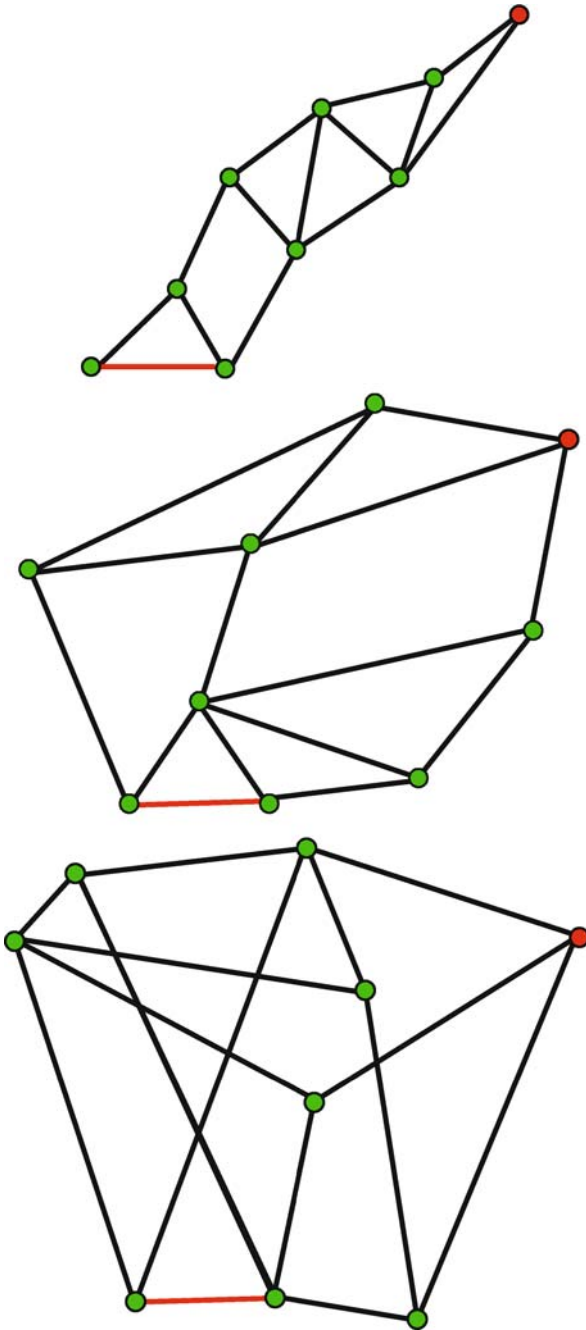
As well as developing the famous equations that describe electromagnetism, Maxwell also studied the stability of structures [1,9], although this work is very much less well known. He represented a structure, for example a trestle bridge, by a series of vertices and edges, and then estimated whether the structure was stable or not. In two dimensions, a triangle is rigid but a quadrilateral can be deformed as there is one internal degree of freedom. In general the number of degrees of freedom  $F$  is given by

$$F = dN - N_c$$

where there are  $N$  points embedded in a  $d$  dimensional space with  $N_c$  constraints, where a constraint is an edge. Examples are given in Fig. 1

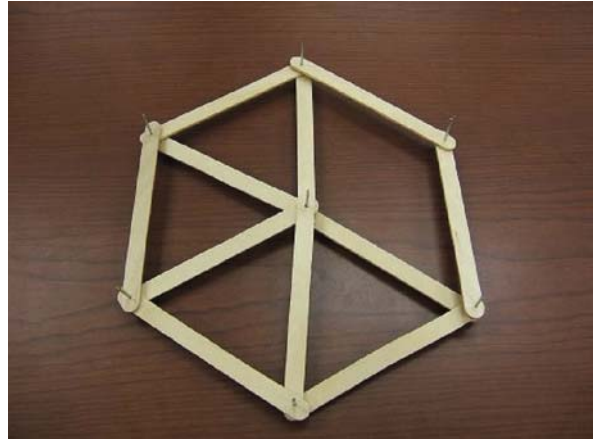
As an introduction, we use popsicle sticks and cotter pins (also called split pins) to construct framework structures. Each popsicle stick has a small hole drilled near each end, through which the cotter pins can be placed. The pin put through two or more sticks, as shown in Fig. 1, in order to construct a two dimensional framework. A pin allows for possible hinge type motion at the joint, even if the two parts of the split pin are opened. A triangle of popsicle sticks is rigid as no internal deformations are possible, and in a similar way the two edge-sharing popsicles in the upper part of Fig. 2 are rigid.

An individual popsicle stick on a two dimensional surface has 3 degrees of freedom. This can be most easily seen as the two degrees of freedom associated with the center of mass plus a rotation about the center of mass. Thus if there are a total of  $N_{st}$  popsicle sticks, then there are initially  $3N_{st}$  degrees of freedom (replacing  $dN$  in the previous equation) before any cotter pins are added to form joints. When a cotter pin is placed through the holes in two sticks, the holes must be aligned in both the  $x$  and  $y$  directions, giving two constraints. Each additional



**Networks, Flexibility and Mobility in, Figure 1**  
 Showing three examples of two-dimensional isostatic networks consisting of 9 vertices and 15 edges

popsicle stick added onto a cotter pin, means two additional constraints, so that the total number of constraints is  $N_c = 2 \sum_r n_r(r-1)$ , where there are  $n_r$  joints with  $r$  popsicle sticks coming together. Thus the number of de-



**Networks, Flexibility and Mobility in, Figure 2**  
 Showing how popsicle sticks and cotter pins can be used to construct framework structures

grees of freedom can be written as

$$F = 3N_{st} - 2 \sum_r n_r(r-1)$$

where it is convenient to let the sum include dangling ends where  $r = 1$  in what follows. The total number of sticks is  $N_{st} = \sum_r n_r r/2$  and the total number of cotter pins is  $N_{cp} = \sum_r n_r$  so that the equation above can be written in the convenient form

$$F = 2N_{cp} - N_{st}.$$

From Fig. 2, we have  $N_{st} = 11$  and  $N_{cp} = 7$ , so that from the equation above, the number of floppy modes  $F = 3$ . These 3 are the macroscopic motions of the whole structure (two translations and one rotation) and there are no internal floppy modes. Note that the surprisingly simple formula above applies even if frameworks with unattached popsicle sticks (dangling ends with  $r = 1$ ) are present as long as a cotter pin is placed in the hole of the dangling end. Although such a pin serves no purpose, it does help simplify the counting and leads to the simple equation above.

The approach above needs correction when redundant popsicle sticks are present – for example if the last spoke is included in Fig. 2, then there is a single redundant popsicle stick present and the count needs modifying. Note that removing any one of the 12 popsicle sticks in such a framework will remove the redundancy. The great challenge of rigidity theory is going beyond the simple counts given here to account for redundancy.

The approach taken here focused on the popsicle sticks as the fundamental objects or bodies, and the cotter pins

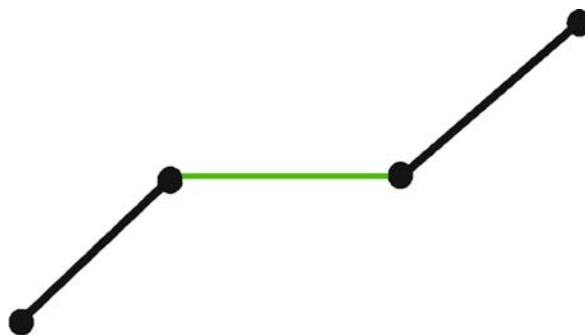
as constraints. However it is often convenient to reverse this and treat the cotter pins as point objects and the popsicle sticks as constraints. The meaning of the equation above now becomes clearer. Each cotter pin has 2 degrees of freedom, as it is a point in a two dimensional plane and so no rotations are involved. This is the origin of the term  $2N_{cp}$  and the term  $N_{st}$  represents the number of constraints ( $N_c$ ). Hence in further analysis in two dimensions in this section, we will use the first equation in this section with  $d = 2$  to give

$$F = 2N - N_c.$$

An example is given in Fig. 1, where each of the three panels has  $d = 2$ ,  $N = 9$  and  $N_c = 15$ , so that  $F = 3$ . These three degrees of freedom are the two translations and the one rotation, associated with the rigid body macroscopic motions of the whole system. This can be thought of in terms of surveying, where the conventional technique using triangulation produces a graph that is made up of a series of edge sharing triangles as shown in the upper left panel. The baseline, together with the distant point to be located with respect to the baseline, are shown in red. This produces an isostatic graph as do the other two panels, although these would probably not be used by any practicing surveyor! Thus, there are no internal motions within these three isostatic networks. This kind of single global count is referred to as the Maxwell count, and is accurate as long as all the counted constraints are independent. The problem of course is that it is hard to know whether constraints are independent or not a priori, especially in very large graphs. Thus, the Maxwell count should be thought of as an estimate or global count, which is good if the network is fairly homogeneous as in the three examples in Fig. 1.

Adding additional edges to the graphs in Fig. 1, leads to redundancy, which is of course very desirable in a building, where one does not want to have the building collapse if a single beam (edge) fails. It also leads to stability against mutations and thermal motions in proteins [10]. If the number of edges associated with each vertex varies a lot throughout the system, then the Maxwell count is incorrect and may or may not be useful as an estimate. In the examples in Fig. 1, the Maxwell count is exact, and there are just enough connections to hold the system rigid, making all three graphs isostatic. If an edge is removed in any of three graphs, there will be a single internal degree of freedom or floppy mode and the graph is called hypostatic. If an additional edge is added then there is a single redundant edge and the graph is called hyperstatic.

The study of such graphs to determine flexibility does not involve any motion, and so belongs in the realm of



**Networks, Flexibility and Mobility in, Figure 3**

Showing a molecular hinge in three dimensions where dihedral angle rotations are allowed using the *green bond* as an axis

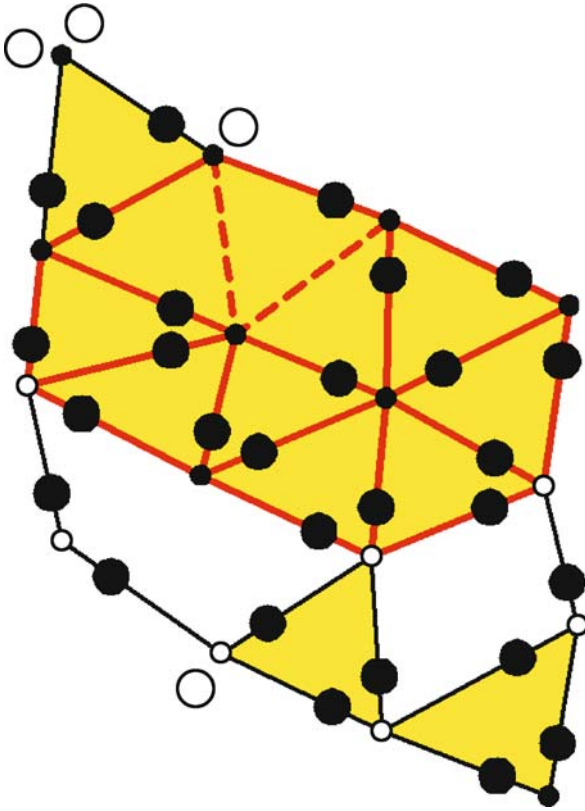
statics. That is the knowledge that a region is flexible, determines the potential for motion, but does not determine its amplitude. Here we are dealing with virtual displacements. To study mobility, it is necessary to make actual displacements and study the motion of the graph. Here an additional set of inequality constraints may be introduced, such that each vertex becomes a hard disc (in two dimensions) or a hard sphere in (three dimensions) where such objects are not allowed to interpenetrate. Such inequality constraints are irrelevant for the static analysis of a graph, but greatly cut down on the mobility of the flexible regions when motion is allowed.

Macromolecules and disordered solids can be represented by graphs, where the atoms are the vertices and the distance between nearest neighbor atoms are the edges. This is particularly appropriate for covalent bonds [11], where the bond length is fixed to within a few percent and the bond angle between adjacent covalent bonds that share a common atom is also fixed similarly and can be represented by an edge connecting second neighbors. The system is still potentially flexible in three dimensions as dihedral angle rotations are allowed, involving a twist along a linear section with four atoms, if no other constraints are present to forbid such motion. This is a three-dimensional example of a hinge and is shown in Fig. 3.

## Flexibility

A graph is defined by its vertices and edges, which can be used to represent a physical system like a glass or a protein, where the vertices are atoms. Laman's theorem [3] in two dimensions can be thought of as a Maxwell count but applied on all length scales to each subgraph. This is a topological theorem. The Maxwell count is also topological, in that the particular value of the lengths of the edges are irrelevant. Such graphs are called generic and have no special angles or lengths associated with them, and represent





**Networks, Flexibility and Mobility in, Figure 4**

Showing the result of a pebble game in two dimensions, where the four free pebbles (*unshaded*) designate the three macroscopic rigid body modes (two translations and one rotation) and the single internal floppy mode, which can be visualized as a rocking of the flower part of the graph

the whole class of such graphs. Graphs that do have symmetries like parallel lines, are called atypical and there is no general theory of the flexibility of such graphs although much is known as a result of work by Dove [12,13,14,15] and Guest [16,17]. This is a rare case where the existence of symmetry is not a simplification, but rather a serious complication that can introduce additional degrees of freedom. Alternative methods using group theory appropriate for the symmetry can be used in crystalline lattices [18] and icosahedral viral capsids [16].

Laman's theorem [3] has been put into an algorithmic form by Hendrickson [19] and implemented by Jacobs and Thorpe [20,21,22] in an algorithm called the pebble game. In this algorithm, two free pebbles are associated with each vertex in the graph and can be moved onto independent edges in order to maintain the proper book keeping between degrees of freedom and constraints. An example of the output of a pebble game search for a small graph is shown in Fig. 4.

### Maxwell Counting

We return now to the global count as first introduced by Maxwell. This is a very simple and useful way to estimate if a network is rigid. Let us consider a regular lattice like a triangular net in two dimensions and a face centered cubic lattice in three dimensions. These must be slightly distorted to make them generic. If we have  $N$  sites, each with  $z$  neighbors present independently with probability  $p$ , then there are  $dN$  degrees of freedom and  $Nzp/2$  constraints (assumed independent) associated with the bonds between nearest neighbors. Thus, the Maxwell estimate of the number of degrees of freedom is given by

$$F = dN - \frac{Nzp}{2} = dN \left( 1 - \frac{\langle r \rangle}{2d} \right)$$

where we define a mean coordination  $\langle r \rangle = zp$ . This means that the number of floppy modes goes to zero at

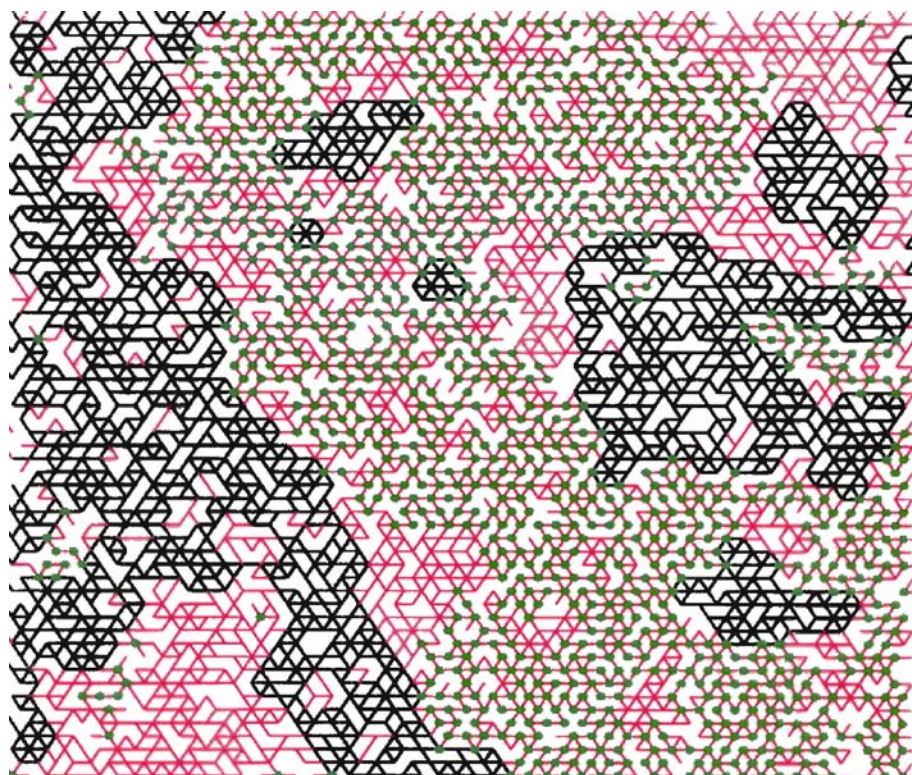
$$\langle r \rangle = zp_c = 2d$$

which gives an estimate for the critical fraction of bonds present at rigidity percolation at  $p_c = 2d/z$ . Of course the Maxwell count fails as the phase transition is approached, and fails catastrophically above the phase transition, where it yields a negative number of floppy modes! This estimate gives  $p_c = 2/3$  for the triangular network in two dimensions and  $p_c = 1/2$  for the face centered cubic lattice in three dimensions. These estimates are very close to the best current numerical estimates determined using the pebble game of 0.6602 [21] and 0.4967 [5], respectively. This shows the remarkable accuracy of the Maxwell estimates for rigidity percolation. It is sometimes said that the Maxwell count is a mean field theory, but we prefer to use the term global count [21]. Nevertheless, there are hidden long range effects in rigidity which is reminiscent of mean field theory [23]. Note that the phase transition is second order in two dimensions [21] but can be first order in three dimensions [5], but nevertheless the Maxwell estimates  $p_c$  are excellent in both cases. A typical sample of a triangular net very near percolation is shown in Fig. 5.

### Enumeration Methods

Although the Maxwell count gives a good global picture in many cases, it gives no information about site to site variations. This was accomplished by Laman [3] in 1970 who showed how the Maxwell count should be applied to each and every subgraph of the network to get a complete microscopic solution as shown for example in Fig. 5. Laman's theorem was turned into an algorithm by Hendrickson [19] that has been implemented on the computer by Jacobs and





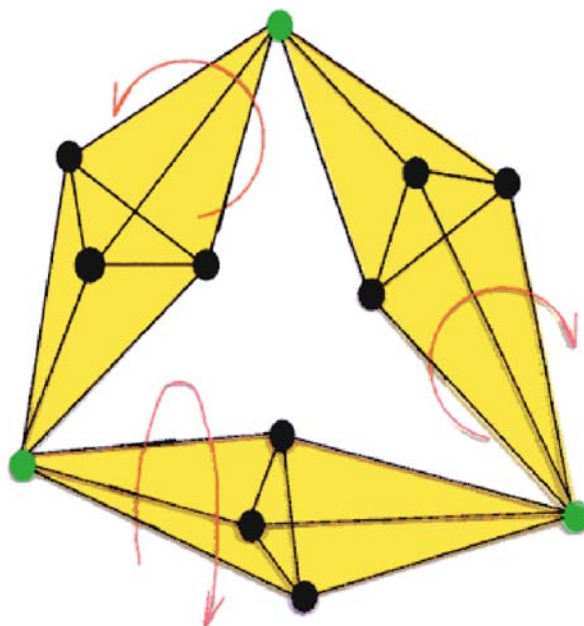
**Networks, Flexibility and Mobility in, Figure 5**

Showing a piece of a triangular network very near percolation, caused by bond dilution. Hinges are shown in *green*, isostatic regions in *red* and hyperstatic regions in *black*

Thorpe [20,21]. This leads to the rigid region decomposition of the kind shown in Fig. 5. This ability to examine very large networks with 100,000 atoms and more and to obtain a rather complete description of the second-order phase transition for the triangular net and the associated geometrical critical exponents [21,24] is extremely useful. Response functions, like elastic constants can also be calculated, but much less accurately as more traditional approaches like conjugate gradient methods must be used on smaller lattices [25]. The elastic constants go to zero as the phase transition is approached from the rigid side for bond dilution on both the triangular lattice and the face-centered cubic lattice [25] which is surprising as the phase transition for the face-centered cubic lattice is first order. This result requires further explanation.

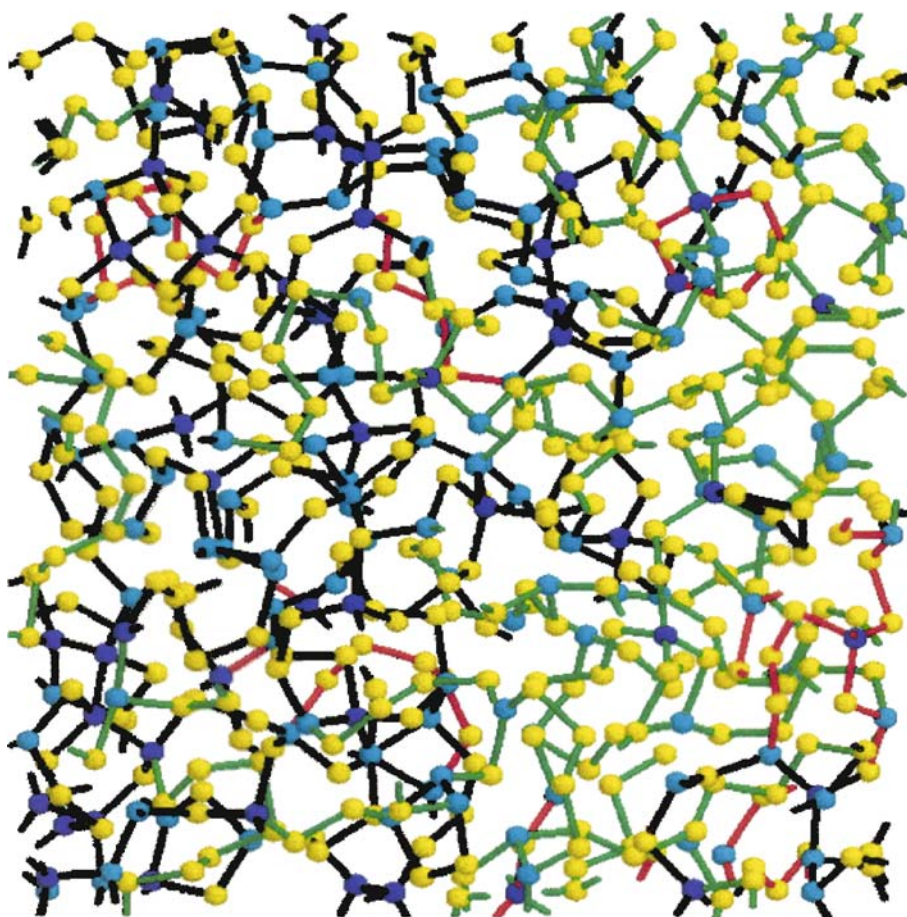
### Molecular Framework Conjecture

Rigidity in three dimensions is much more complicated than in two dimensions because of the existence of “banana diagrams”, an example of which is shown in Fig. 6. The three rigid yellow pieces are the bananas and it



**Networks, Flexibility and Mobility in, Figure 6**

Showing the problematic banana diagram in three dimensions



Networks, Flexibility and Mobility in, Figure 7

Showing a piece of a covalent glass near the phase transition at a mean coordination  $\langle r \rangle = 2.4$ , where the *green bonds* are hinges, the *red bonds* isostatic and the *black bonds* hyperstatic

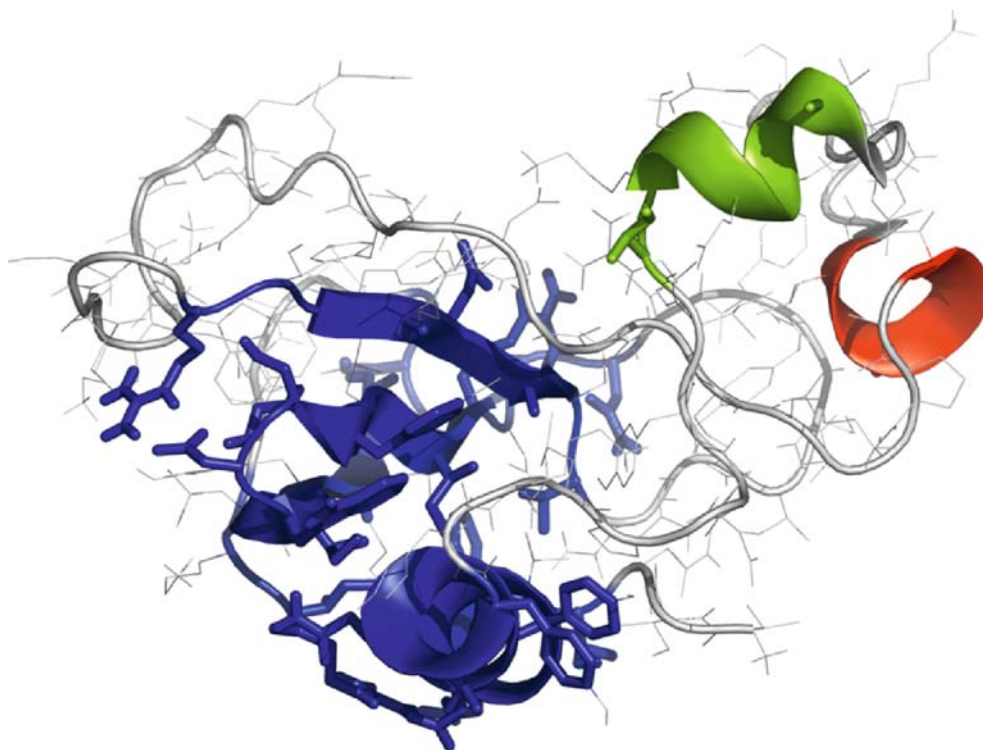
can be seen that these can rotate as indicated by the arrows. The complication is that the three green sites themselves form a non-contiguous rigid cluster. This cannot happen in two dimensions (see Fig. 4 and 5), where all the individual rigid clusters are contiguous and represent a major complication. It means that there is no Laman-type theorem and hence no pebble game. Nevertheless, a very accurate, but approximate, pebble game has been developed for general three-dimensional networks [5] that for example is estimated to be accurate to 1 part in  $10^9$  for bond diluted face-centered cubic lattices.

There is one very important special case in three dimensions, which is the bond-bending network, where there are angular forces associated every each pair of central forces that share a common vertex. These angular forces can be thought of as second neighbor central forces. For these special networks there is a conjecture; called

the *Molecular Framework Conjecture* that the appropriate pebble game is exact [4,26,27]. This is a very fortunate situation as this is the case of most practical interest. Examples of such molecular frameworks are covalent glasses like GeAsSe as shown in Fig. 7, where the coloring of the sites reflects the chemistry (dark blue for 4-coordinated Ge, light blue for 3-coordinated As and yellow for 20-coordinated Se), and the bond coloring reflects whether they are flexible (green), rigid (red) or stressed and rigid (black).

Maxwell counting can be used to estimate when these networks undergo a transition from rigid to flexible. If there are  $n_r$  sites with  $r$ -coordinated atoms, where  $r$  is 2, 3 or 4, then the total number of atoms  $N = \sum_r n_r$  and the mean coordination  $\langle r \rangle = \sum_r r n_r / N$ . The total number of degrees of freedom is  $3N$ . The number of angular constraints is  $r/2$  for central forces and  $3r - 5$  for angular





#### Networks, Flexibility and Mobility in, Figure 8

Showing a piece of the protein barnase with the three largest rigid regions picked out in the three different solid colors

forces, both associated with an  $r$ -coordinated site. Therefore, the number of floppy modes is

$$F = 3N - \sum_r n_r \left[ \frac{r}{2} + (2r - 3) \right] = 6N \left[ 1 - 5 \frac{\langle r \rangle}{12} \right]$$

which gives a phase transition from rigid to flexible when  $\langle r \rangle = 12/5 = 2.4$  [28,29,30]. Note that reference [28] which first applied these counting ideas to network glasses, contains an error as six angular constraints rather than the correct five independent angular constraints were assigned to a 4-coordinated site. Exact enumerations, using the pebble game on computer generated networks give  $\langle r \rangle = 2.385$  for the location of the phase transition [31]. Thus, when the polymer chains made up of 2-coordinated atoms are lightly cross-linked, the network is flexible, but as the number of cross-links becomes quite dense and the mean coordination rises to around 2.4, the network becomes rigid [29].

The molecular framework conjecture also applies to proteins, which are large macromolecules containing many hundreds of atoms. Proteins are polypeptide chains that are crossed linked by hydrophobic tethers and hydrogen bonds to form rather compact three-dimensional

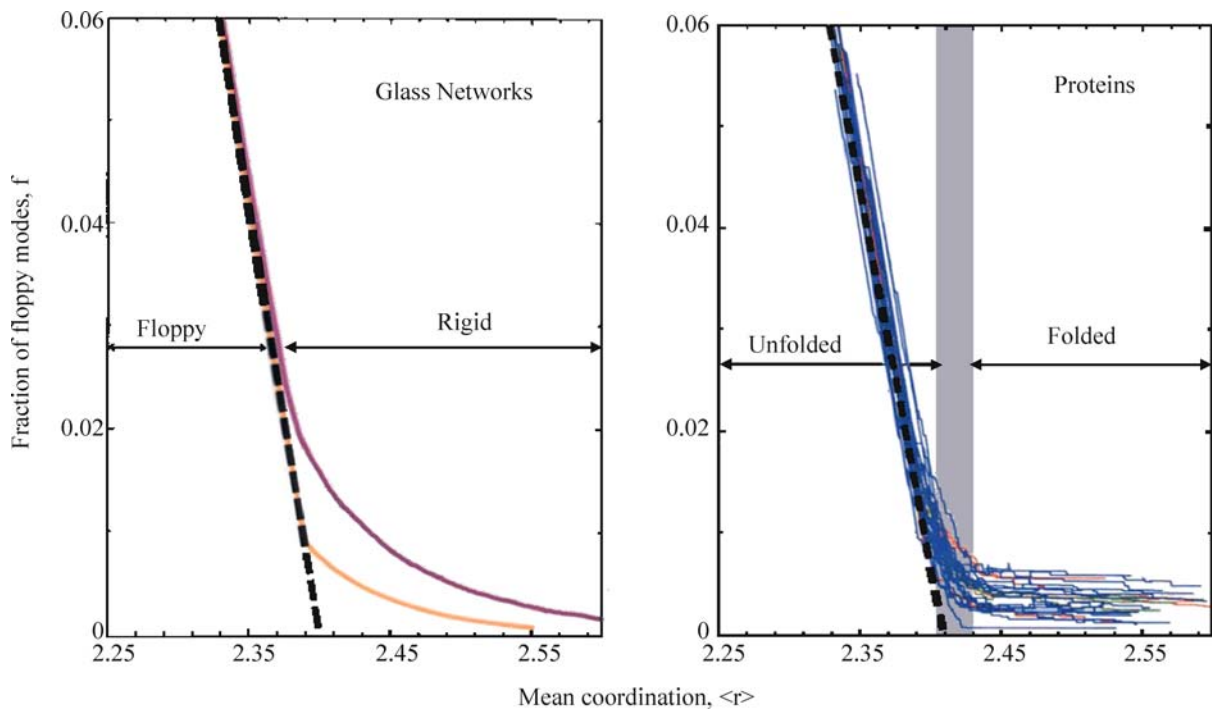
structures [32]. A rigid region decomposition of a typical protein, barnase is shown in Fig. 8.

Proteins have enough rigidity to define their three-dimensional structure, while retaining sufficient flexibility to function. Thus, they exist in a narrow window around the phase transition between flexible and rigid [10]. This is shown in Fig. 9, which also shows that covalent glasses behave in a very similar way.

#### Geometrical Simulation

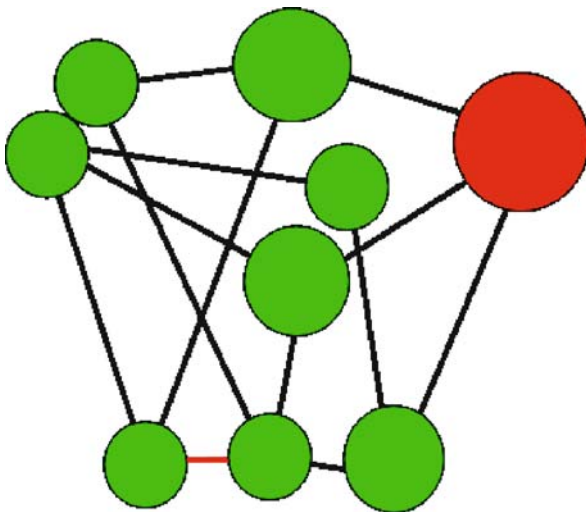
The approach described here comes from an area of mathematics called rigidity theory, which yields powerful results without much calculational or computational overhead. However, these results are limited as no actual motion is involved, and it is only the possibility of motion, or absence of motion, that is addressed. This is very valuable information.

However, for a more complete description, it is desirable to know the mobility, or the amplitudes of the motion, which do not violate any of the original constraints. In addition, the motion may involve additional constraints, the most important being inequalities that prevent objects from interpenetrating. For example, two-dimensions



#### Networks, Flexibility and Mobility in, Figure 9

Showing the similar behavior of the number of floppy modes in covalent network glasses and a selection of proteins [10]. In all cases the transition from flexible to rigid occurs around a mean coordination  $\langle r \rangle = 2.4$ . Here  $f = F/3N$  is the number of floppy modes per degree of freedom



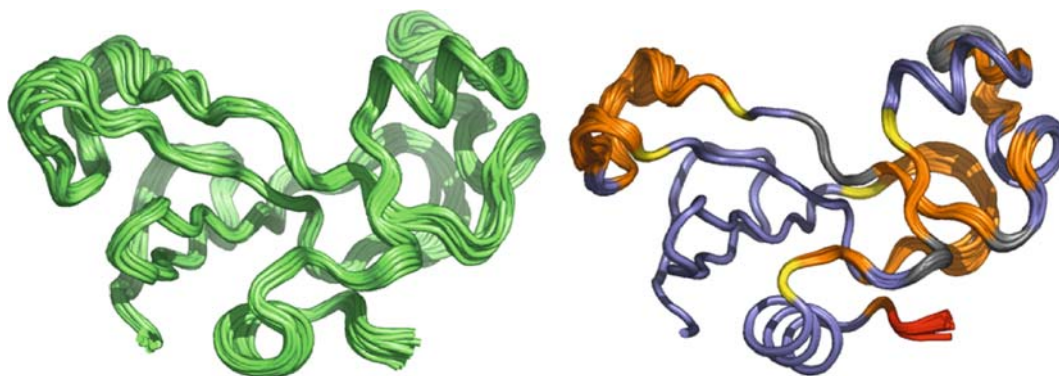
#### Networks, Flexibility and Mobility in, Figure 10

Showing the two-dimensional graph from the lower panel in Fig. 1, with a single edge removed, so that the graph becomes hypostatic with a single floppy mode

discs are not allowed to overlap, and in three dimensions, spheres are prevented from interpenetrating. An example is shown in Fig. 10, where the vertices from the lower panel in Fig. 1 have been replaced by discs of various radii.

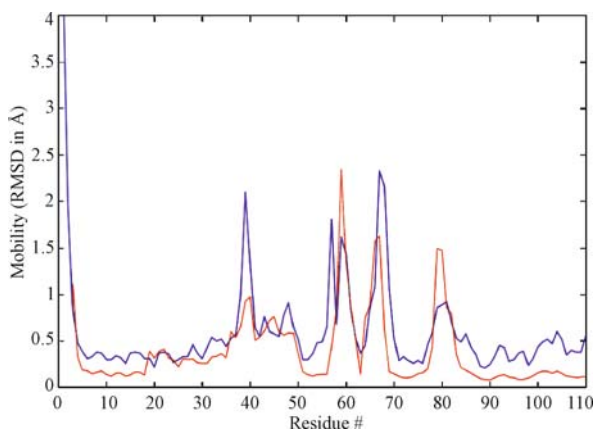
Pairs of discs are not allowed to overlap during the motion. As the disc radii are increased, the amplitude of the floppy mode is liable to be decreased, by the necessity to avoid collisions. Geometric simulation methods can be used to study such motion [33,34,35] and these are rather efficient. Key ingredients are the introduction of random or directed small displacements, the maintenance of constraints and collision avoidance. The reader is referred to the articles above for more details.

Geometrical simulation techniques are similar in three dimensions, and for example new conformations of molecules can be generated. An example is shown in Fig. 11 for the protein barnase, previously shown in Fig. 8. A set of 20 conformers obtained from a nuclear magnetic resonance experiment [36] are shown in the left panel. These are determined from experiments where a sufficient number of constraints are determined that the remaining indeterminacy is associated with the protein mobility. In the right panel, a single X-ray structure [37] is used as a starting point to generate many new protein conformations consistent with the constraints. The 20 simulation conformers shown are selected from a much larger set and are spaced well apart [6]. The blue regions in the left panel are rigid, and the other regions are flexible, leading to mobility. Note that distinct



**Networks, Flexibility and Mobility in, Figure 11**

Showing a comparison between the NMR ensemble of conformers (*left*) for barnase and the set of conformers (*right*) generated using geometrical simulation techniques from a single X-ray crystallographic structure [6]



**Networks, Flexibility and Mobility in, Figure 12**

Plotting the mobility of each residue, as determined using experimental data (*blue*) from Fig. 11, and showing that geometric simulation (*red*) captures the main features of the mobility of the protein, when compared with nuclear magnetic resonance data [6]

rigid regions can move with respect to other rigid regions.

A quantitative comparison between the Root Mean Square Distance (RMSD) and the residue number between the nuclear magnetic resonance experiment and the geometric simulations is shown in Fig. 12. The amplitudes obtained from both the experiment and from geometrical simulation are absolute, so no scaling is required. The residue number refers to the numbering of the amino acids along the backbone of the polypeptide chain that folds to produce the three-dimensional protein structure.

### Future Directions

In this brief article, we have set down the key ingredients that are needed when considering the flexibility and

mobility of networks, with an emphasis on the important class of three-dimensional molecular networks. For more details of the more mathematical aspects of the theory of rigidity, the reader is referred first to the books referenced at the end. There are few review articles in this area as these approaches are quite new and are still being developed.

The static part of rigidity has led to an exact theorem [3] and fast algorithms in two dimensions [20,21,22]. There is no theorem in three dimensions for the general case, although there is the special case of the very useful molecular framework conjecture [4] that remains to be proved. Algorithms have been developed recently based on the molecular framework conjecture [31,32,38,39,40]. Although there is no theorem or conjecture for the general case in three dimensions, approximate, but rather accurate, algorithms nevertheless have been developed [5]. There is a need for a more complete description of the unique features of rigidity in three dimensions, associated with banana diagrams etc. At this time there seems no great motivation for studying rigidity in dimensions higher than three.

Geometrical simulation is still in its infancy – this is a more complex problem than the movement of robot arms [41] because many interlocking rings of constraints are involved and the environment is crowded in general. This is rather like moving groups of friends holding hands tightly in rings through a football crowd at the end of a game. Early efforts focused on ring closure [33,42], but these proved inefficient and have been largely abandoned in favor of small Monte Carlo-type moves that initially violate the constraints, followed by various guiding and shaking procedures to restore the constraints and arrive at a new conformation [6]. The inequalities associated with the van der Waals or hard sphere overlaps cause particular concern in the crowded molecular environments



found in proteins and better ways of handling this are needed. Ultimately the results of such approaches, obtained rapidly, can be used as input with phenomenological classical potentials [43] of the kind used in molecular dynamics [44,45], to be able to generate larger amplitude motions for a given amount of computer time.

## Bibliography

### Primary Literature

- Maxwell JC (1864) On the calculation of the equilibrium and stiffness of frames. *Philos Mag* 27:294–299
- Lagrange J-L (1788) *Mécanique Analytique*, 4th edn. Gauthier-Villars, Paris
- Laman G (1970) On graphs and rigidity of plane skeletal structures. *J Eng Math* 4:331–340
- Whiteley W (2005) Counting out to the flexibility of molecules. *Phys Biol* 2:S116–26
- Chubynsky MV, Thorpe MF (2007) Algorithms for 3d rigidity analysis and a first order phase transition. *Phys Rev E* 76:041135
- Wells SA, Menor S, Hespenheide BM, Thorpe M (2005) Constrained geometric simulation of diffusive motion in proteins. *Phys Biol* 2:127–136
- Goldstein H, Poole CP, Safko JL (2002) *Classical mechanics*, 3rd edn. Addison Wesley, San Francisco, p 638
- Wales D (2003) *Energy landscapes*. Cambridge University Press, Cambridge
- Maxwell JC (1864) On reciprocal figures and diagrams of forces. *Phil Mag* 27:250–261
- Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: Rigidity lost. *Proc Natl Acad Sci USA* 99:3540–5
- Pauling L (1928) The shared-electron chemical bond. *Proc Natl Acad Sci USA* 14:359–62
- Dove MT, Heine V, Hammonds KD (1995) Rigid unit modes in framework silicates. *Mineral Mag* 59:629–639
- Dove MT, Hammonds KD, Heine V, Withers RL, Kirkpatrick RJ (1996) Rigid unit modes in the high-temperature phase of  $\text{SiO}_2$  tridymite: Calculations and electron diffraction. *Phys Chem Miner* 23:56–62
- Dove MT, Keen DA, Hannon AC, Swinson IP (1997) Direct measurement of the Si–O bond length and orientational disorder in the high-temperature phase of cristobalite. *Phys Chem Miner* 24:311–317
- Dove MT, Trachenko KO, Tucker MG, Keen DA (2000) Rigid unit modes in framework structures: Theory, experiment and applications. *Transform Process Miner* 39:1–33
- Kangwai RD, Guest S (1999) Detection of finite mechanisms in symmetric structures. *Int J Solid Struct* 36:5507–5527
- Fowler PD, Guest S (2002) Symmetry and states of self-stress in toroidal frames. *Int J Solid Struct* 39:4385–4393
- Ashcroft NW, Mermin ND (1976) *Solid state physics*. Holt Rinehart and Winston, New York, p 826
- Hendrickson B (1992) Conditions for unique graph realizations. *Siam J Comput* 21:65–84
- Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: The pebble game. *Phys Rev Lett* 75:4051–4054
- Jacobs DJ, Thorpe MF (1996) Generic rigidity percolation in two dimensions. *Phys Rev E* 53:3682–3693
- Jacobs D, Hendrickson B (1997) An algorithm for two-dimensional rigidity percolation: The pebble game. *J Comp Phys* 137:346–365
- Kittel C, Shore H (1965) Development of a phase transition for a rigorously solvable many-body system. *Phys Rev* 138:A1165–A1169
- Arbabi S, Sahimi M (1993) Mechanics of disordered solids. I Percolation on elastic networks with central forces. *Phys Rev B Condens Matter* 47:695–702
- Feng S, Thorpe MF, Garboczi E (1985) Effective-medium theory of percolation on central-force elastic networks. *Phys Rev B* 31:276–280
- Whiteley W (1999) Rigidity of molecular structures: Generic and geometric analysis. In: Thorpe MF, Duxbury PM (eds) *Rigidity Theory and Applications*. Kluwer Academic/Plenum Publishers, New York
- Tay T-S, Whiteley W (2005) Comparison of molecular models. unpublished
- Phillips J (1979) Topology of covalent non-crystalline solids. 1 Short-range order in chalcogenide alloys. *J Non-Cryst Solid* 34:153–181
- Thorpe MF (1983) Continuous deformations in random networks. *J Non-Cryst Solid* 57:355–370
- He H, Thorpe MF (1985) Elastic properties of glasses. *Phys Rev Lett* 54:2107–2110
- Thorpe MF, Jacobs DJ, Chubynsky NV, Rader AJ (1999) Generic rigidity of network glasses. In: Thorpe MF, Duxbury PM (eds) *Rigidity theory and applications*. Kluwer Academic/Plenum Publishers, New York, pp 239–277
- Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. *Proteins* 44:150–65
- Lei M, Zavodszky MI, Kuhn LA, Thorpe MF (2004) Sampling protein conformations and pathways. *J Comput Chem* 25:1133–48
- Lee A, Streinu I, Brock O (2005) A methodology for efficiently sampling the conformation space of molecular structures. *Phys Biol* 2:S108–S115
- Wells S, Menor S, Hespenheide B, Thorpe MF (2005) Constrained geometric simulation of diffusive motion in proteins. *Phys Biol* 2:S127–36
- Bycroft M, Ludvigsen S, Fersht AR, Poulsen FM (1991) Determination of the three-dimensional solution structure of barnase using nuclear magnetic resonance spectroscopy. *Biochemistry* 30:8697–701
- Hartley RW (1989) Barnase and barstar: Two small proteins to fold and fit together. *Trends Biochem Sci* 14:450–4
- Jacobs D (1998) Generic rigidity in three-dimensional bond-bending networks. *J Phys A Math Gen* 31:6653–6668
- Jacobs D, Kuhn LA, Thorpe MF (1999) Flexible and rigid regions in proteins. In: Thorpe MF, Duxbury PM (eds) *Rigidity Theory and Applications*. Kluwer Academic/Plenum Publishers, Traverse City
- Hespenheide BM, Jacobs DJ, Thorpe MF (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. *J Phys Condens Matter* 16:S5055–S5064
- Streinu I (2000) A combinatorial approach to planar non-colliding robot arm motion planning. In: *Proc 41st Annual ACM/IEEE Symposium on Foundations of Computer Science (FOCS 2000)*. IEEE Computer Society Washington, DC, pp 443
- Kuhn LA, Zavodszky MI, Arora S, Lei M, Thorpe MF (2004) Modeling correlated protein main-chain and side-chain motions

in ligand docking and screening. Abstr Pap Amer Chem Soc 228:U501

43. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. *Nature* 267:585–90
44. Brooks RR, Bruccoleri BE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217
45. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, Debolt S, Ferguson D, Seibel G, Kollman P (1995) Amber, a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. *Comput Phys Commun* 91:1–41

## Books and Reviews

- Boolchand P (2000) *Insulating and Semiconducting Glasses*. Series on Directions in Condensed Matter Phys 17. World Scientific, Singapore
- Graver JE, Servatius B, Servatius H (1993) *Combinatorial Rigidity*. In: Graduate Series in Mathematics. Amer Math Soc
- Thorpe MF, Duxbury PM (1999) *Rigidity Theory and Applications*. Kluwer Academic, Plenum Press, New York
- Thorpe MF, Tichy L (2001) Properties and Applications of Amorphous Materials. In: Thorpe MF, Tichy L (eds) *NATO Science Series, II Mathematics, Physics and Chemistry*, vol 9

## Networks and Stability

FRANK H. PAGE JR.<sup>1,2</sup>, MYRNA WOODERS<sup>3,4</sup>

<sup>1</sup> Department of Economics, Indiana University, Bloomington, USA

<sup>2</sup> Centre d'Economie de la Sorbonne, Université Paris 1, Pantheon-Sorbonne, France

<sup>3</sup> Department of Economics, Vanderbilt University, Nashville, USA

<sup>4</sup> Department of Economics, University of Warwick, Coventry, UK

## Article Outline

Glossary

Definition of the Subject

Introduction

The Primitives

Abstract Games of Network Formation and Stability

Strong Stability, Pairwise Stability, Nash Stability, and Farsighted Consistency

Singleton Basins of Attraction

Future Directions

Acknowledgments

Bibliography

## Glossary

**Homogeneous networks** A homogeneous network consists of a finite set of nodes together with a finite set of mathematical objects called links or arcs, each identifying a connection between a pair of nodes. Given finite node set  $N$  with typical element  $i$ , a homogeneous linking network  $G$  is a finite collection of sets of the form  $\{i, i'\}$  called links. Link  $\{i, i'\} \in G$  indicates that nodes  $i$  and  $i'$  are connected in network  $G$ . A homogeneous directed network  $G$  is a finite collection of ordered pairs  $(i, i')$  called arcs. Arc  $(i, i') \in G$  indicates that nodes  $i$  and  $i'$  are connected in network  $G$  via a connection running from  $i$  to  $i'$ . In a homogeneous network (whether it be a linking network or a directed network) all connections are of the same type.

**Heterogeneous networks** A heterogeneous network consists of a finite set of nodes together with a finite set of mathematical objects called labeled links or labeled arcs, each identifying a particular type of connection between a pair of nodes. Given finite node set  $N$  with typical element  $i$  and given finite label set  $A$  with typical element  $a$ , a heterogeneous linking network  $G$  is a finite collection of ordered pairs of the form  $(a, \{i, i'\})$  called labeled links. Labeled link  $(a, \{i, i'\}) \in G$  indicates that nodes  $i$  and  $i'$  are connected in network  $G$  via a type  $a$  link. A heterogeneous directed network  $G$  is a finite collection of ordered pairs of the form  $(a, (i, i'))$  called labeled arcs. Labeled arc  $(a, (i, i')) \in G$  indicates that nodes  $i$  and  $i'$  are connected in network  $G$  via a type  $a$  arc running from  $i$  to  $i'$ . In a heterogeneous network (whether it be a linking network or a directed network) connections can differ and are distinguished by type.

**Abstract game of network formation with respect to irreflexive dominance** An abstract game of network formation with respect to irreflexive dominance consists of a feasible set of networks  $\mathbb{G}$  equipped with a irreflexive dominance relation  $>$ . A dominance relation on  $\mathbb{G}$  is a binary relation on  $\mathbb{G}$  such that for all  $G$  and  $G'$  in  $\mathbb{G}$ ,  $G' > G$  (read  $G'$  dominates  $G$ ) is either true or false. The dominance relation is irreflexive if  $G > G$  is always false.

**Abstract game of network formation with respect to path dominance** An abstract game of network formation with respect to path dominance consists of a feasible set of networks  $\mathbb{G}$  equipped with a path dominance relation  $\geq_p$  induced by an irreflexive dominance relation  $>$  on  $\mathbb{G}$ . Given networks  $G$  and  $G'$  in  $\mathbb{G}$ ,  $G' \geq_p G$  (read  $G'$  path dominates  $G$ ) if either  $G' = G$  or there

is a finite sequence of networks in  $\mathbb{G}$  beginning with  $G$  and ending with  $G'$  such that each network along the sequence dominates its predecessor.

### Definition of the Subject

Our subject is networks, and in particular, stable networks and the game theoretic underpinnings of stable networks.

Networks are pervasive. We routinely communicate over the internet, advance our careers by networking, travel to conferences over the transportation network and pay for the trip using the banking network. Doing this utilizes networks in our brain. The list could go on. While network models have had a long history in sociology, the natural sciences, and engineering (e.g., in modeling social organizations, brain architecture, and electrical circuits), the rise of the network paradigm in economics is relatively recent. Economists are now beginning to think of political and economic interactions as network phenomena and to model everything from terrorist activities to asset market micro structures as *games of network formation*. This trend in economics, which began with the seminal paper by Myerson [88] on graphs and cooperation and accelerated with the publication of the papers by Jackson and Wolinsky [64] and Dutta and Mutuswami [39] on stable and efficient networks, is likely to continue with the development of new algorithms, the expansion of computational capacity and the broad application of network theories to economic, political, and social phenomena.

What economists bring to the study of networks that is new is game theory. For the most part sociologists, natural scientists and engineers have used networks descriptively and have focused on the design of networks from the perspective of a single designer or on the random evolution of networks from the perspective of nature. This singularity of perspective is a consequence of the nonstrategic nature of the phenomena being explained or the problem being solved (e.g., the spread of a disease through a given population, the transmission of electrical impulses in the brain, or the optimal design of an integrated circuit). In economics the perspective is often times strategic. In particular, in many economic situations, several individuals, guided by their own self interest, behave strategically in putting into place pieces of the network of economic, political, or social interactions under their control and in so doing generate payoffs and externalities that determine the network of economic interactions that eventually emerges in equilibrium. Thus in economics, pieces of the network are the strategies and the network that ultimately prevails is the result of strategic competition rather than the design of a single individual or nature. Because

large computer networks such as the internet are built, operated, and used by a many diverse individuals with competing interests, computer scientists are also beginning to use game theoretic models to analyze and understand the optimal design of secure computer networks (see for example, [105,115]). Conversely, what networks bring to the study of economics is a new way of modeling the structure of economic interactions and externalities that makes possible a game-theoretic analysis of how these structures influence individual payoffs and the economic equilibrium that emerges from competition.

### Introduction

Our main objective is to present a unified, game-theoretic development of the main concepts of stability that have appeared in the recent economics literature on strategic network formation, specifically, the notions of strong stability (Jackson and van den Nouweland [61]), pairwise stability (Jackson and Wolinsky [64]), Nash stability, and farsighted consistency (Chwe [26]). In order to accomplish this we follow the approach introduced in Page and Wooders [91,93]. The key ingredient in this approach is an abstract game model of endogenous network formation (i.e., abstract game in the sense of von Neumann–Morgenstern [122]). The model is built on four primitives: A feasible set of networks, player preferences over networks, the rules of network formation, and an irreflexive dominance relation over networks. In the remainder of this introduction we provide an overview of the four primitives of our model, a summary of results discussed, and note some important areas of research that are beyond the scope of this entry.

*Feasible Sets:* The feasible set may consist of networks as simple as homogenous linking networks or as complex as heterogeneous directed networks. All networks consist of a finite set of nodes (representing, for example, economic agents or players) together with a finite set of mathematical objects called links, labeled links, arcs, or labeled arcs describing the connections between nodes. Here we will focus on homogeneous linking networks, as does most of literature (see, for example, Myerson [88], Jackson and Wolinsky [64], and Jackson and van den Nouweland [61]), except in our discussion of Nash stability where we will consider homogeneous directed networks (as in [5]). What distinguishes homogeneous networks (linking or directed) from heterogeneous networks (linking or directed) is that in a homogeneous network all connections between nodes are of the same type whether represented by a link as in a linking network or by an arc as in a directed network. Thus, in a homogeneous link-

ing network all links are of the same type and in a homogeneous directed network all arcs are of the same type. While homogeneous networks are quite restrictive, they have been very important in developing our understanding of social and economic networks and have proved very useful in many economic applications (see, for example, Belleflamme and Bloch [7], Bramoulle and Kranton [17], Calvo-Armengol [19], and Furusawa and Konishi [41]). Page and Wooders [89,91,93,94] extend the existing literature on economic and social networks by introducing the notion of heterogeneous directed networks. These types of networks potentially have a rich set applications (in the natural sciences, engineering, sociology, politics, as well as economics) because connections or interactions between nodes can be distinguished by direction or intent as well as by type, intensity, or purpose.

*Players' Preferences:* We will assume throughout that each player's preferences are given by an irreflexive binary relation defined on the feasible set of networks. Thus, we will assume that players have strong (or strict) preferences over networks. Under strong preferences, if a player prefers one network to another, then the player's preference is strict. However, we will comment where appropriate on weak preferences. Under weak preferences, if a player prefers one network to another, then the player's preference is either strict or indifferent.

*Rules of Network Formation:* We will focus here on three different sets of rules: Jackson–Wolinsky rules [64], Jackson–van den Nouweland rules [61], and Bala–Goyal rules [5]. In particular, in our discussions of pairwise stable homogeneous linking networks we will assume that the rules of network formation are the Jackson–Wolinsky rules. Under the Jackson–Wolinsky rules the addition of a link is bilateral (i.e., the two players that would be involved in the link must agree to adding the link), the subtraction of a link is unilateral (i.e., at least one player involved in the link must agree to subtract or delete the link), and network changes take place one link at a time (i.e., only one link can be added or subtracted at a time). In our discussion of strongly stable homogeneous linking networks, we will assume that the rules of network formation are the Jackson–van den Nouweland rules. Under the Jackson–van den Nouweland rules link addition is bilateral, link subtraction is unilateral, and in any one play of the game several links can be added and/or subtracted. Thus the Jackson–van den Nouweland rules are the Jackson–Wolinsky rules without the one-link-at-a-time restriction. Finally, in our discussion of Nash homogeneous directed networks we will assume that the rules of network formation are the Bala–Goyal rules. Under the Bala–Goyal rules an arc may be added or sub-

tracted unilaterally by the initiating player involved in the arc and in any one play of the game only network changes brought about by an individual player are allowed. Note that all three of these sets of rules can be described as being uniform across networks. Under uniform rules the rules for changing a network are the same no matter which status quo network is being changed. Page, Wooders and Kamat [94] allow nonuniform rules and introduce a network representation of nonuniform rules.

*Dominance Relations:* Given players' preferences and the rules of network formation we will define a dominance relation over the feasible set of networks that incorporates both players preferences and the rules. Here we will focus on dominance relations that are either direct or indirect. Under direct dominance players are concerned with immediate consequences of their network formation strategies whereas under indirect dominance players are farsighted and consider the eventual consequences of their strategies.

*General Results:* A specification of the primitives induces two types of abstract games over homogeneous networks: (i) a network formation game with respect to the irreflexive dominance relation induced by preferences and rules, and (ii) a network formation game with respect to *path* dominance induced by this irreflexive dominance relation. We will begin by considering the game with respect to irreflexive dominance and present results on the existence of quasi-stable and stable networks. These results provide a network rendition of classical results from graph theory on the existence of quasi-stable sets and stable sets due to Chvatal and Lovasz [25], Berge [8], and Richardson [100]. We will also present a result on the existence and nonemptiness of the set of farsightedly consistent networks. This result is a network rendition of a result due to Chwe [26] for abstract games.

Next we will consider the game over homogeneous networks with respect to path dominance, and we will conclude that the following results hold:

1. Given preferences and the rules governing network formation, the set of homogeneous networks (linking or directed) contains a unique, finite, disjoint collection of nonempty subsets each constituting a *strategic basin of attraction*. These basins of attraction are the absorbing sets of the competitive process of network formation modeled via the game.
2. A stable set of homogeneous networks (in the sense of von Neumann–Morgenstern) with respect to path dominance consists of one network from each basin of attraction.

3. The path dominance core, defined as the set networks having the property that no network in the set is path dominated by any other homogeneous network, consists of one network from each basin of attraction containing a *single* network. Note that the path dominance core is contained in each stable set and is nonempty if and only if there is a basin of attraction containing a single network. As a corollary, we conclude that any homogeneous network contained in the path dominance core is constrained Pareto efficient.
4. From the results above it follows that if the dominance relation is transitive and irreflexive, then the path dominance core is nonempty.

These results are a special cases of results due to Page and Wooders [93].

### Specific Results for Pairwise Stability, Strong Stability, Nash Stability, and Farsighted Consistency

What are the connections between our notions of stability for homogeneous networks (basins of attraction, path dominance stable sets, and path dominance core) and the notions of strong stability [39,61], pairwise stability [64], Nash stability [5], and farsighted consistency [26,94]? From the general results in [91] and [93] for heterogeneous directed networks, we will conclude for the case of homogeneous networks (linking or directed) that, depending on how we specialize the primitives of the model, the path dominance core is equal to the set of strongly stable networks, the set of pairwise stable networks, or the set of Nash networks. In particular, we will conclude that:

- (a) If path dominance is induced by a direct dominance relation, then in the set of homogeneous linking networks the path dominance core is equal to the set of strongly stable networks.
- (b) If, in addition, the rules of network formation are the Jackson–Wolinsky rules, then in the set of homogeneous linking networks the path dominance core is equal to the set of pairwise stable networks.
- (c) If path dominance is induced by a direct dominance relation and if the rules of network formation are the Bala–Goyal rules, then in the set of homogeneous directed networks the path dominance core is equal to the set of Nash networks.

We can then conclude from (3) above that the existence of at least one basin of attraction containing a single network is, depending on how we specialize primitives, both necessary and sufficient for either (i) the existence of a strongly stable network, or (ii) a pairwise stable network, or (iii) a Nash network.

For path dominance induced by an indirect dominance relation, we can conclude from our prior results that for the case of homogeneous linking networks with Jackson–Wolinsky or Jackson–van den Nouweland rules or for the case of homogeneous directed networks with Bala–Goyal rules, each strategic basin of attraction has a non-empty intersection with the largest farsightedly consistent set of networks. This result, together with (2) above, implies that there always exists a path dominance stable set of homogeneous networks contained in the largest farsightedly consistent set. Thus, the path dominance core is contained in the largest consistent set. In light of our results on the path dominance core and stability (both strong and pairwise), we conclude that if path dominance is induced by an indirect dominance relation, then any homogeneous network contained in the path dominance core (i.e., the farsighted core) is not only farsightedly consistent but also strongly stable, as well as pairwise stable. Other papers using indirect dominance (or variations thereof) and farsighted consistency in games (not necessarily network formation games) include Li [76,77], Xue [129,130], Luo [79], Mariotti and Xue [80], Diamantoudi and Xue [36], and Mauleon and Vannetelbosch [84], Bhattacharya [9], Herrings, Mauleon and Vannetelbosch [55].

We remark that solution concepts defined using dominance relations have a long and distinguished history in the literature of game theory. First, consider the von Neuman–Morgenstern stable set (see [122] and [100]). The vN-M stable set is defined with respect to a dominance relation on a set of outcomes and consists of those outcomes that are externally and internally stable with respect to the given dominance relation. Similarly, Gillies [44] defines the core based on a given dominance relation. These solution concepts, with a few exceptions, have typically been applied to models of economies or cooperative games where the notion of dominance is based on what a coalition can achieve using only the resources owned by its members (cf., Aumann [3]) or a given set of utility vectors for each possible coalition (cf., Scarf [106]). Particularly notable exceptions are Schwartz [107], Panzer, Kalai and Schmeidler [66], Kalai and Schmeidler [67], Shenoy [109], Inarra, Kuipers, and Olaizola [58], and van Deemen [119]. Their motivations are in part similar to ours in that they take as given a set of possible choices for players (here consisting of set of networks) and a dominance relation and, based on these, describe a set of possible or likely outcomes called, by Kalai and Schmeidler, the admissible set. While their examples treat direct dominance, their general results have wider applications.

Because our objective here is to provide a unified game theoretic treatment of the main stability notions for net-



work formation games, many topics related to strategic networks are not covered here. For example, we do not discuss the conflict between stability and efficiency which is the main focus of the important papers by Dutta and Mutuswami [39] and Currarini and Morelli [30], and Mutuswami and Winter [87]. Nor do we treat the topic of network formation and cooperative games, the topic of the seminal paper by Myerson [88] and the excellent book by Slikker and van den Nouweland [113] among many other contributions, or the topic of network formation and evolution treated in Hojman and Szeidl [56]. Our game theoretic approach provides a snapshot of all possible network formation paths under dynamics which respect preferences and the rules of network formation. Our approach, however, is not explicitly dynamic. For network dynamics, we can only suggest to the reader the elegant papers by Skyrms and Pemantle [111], Watts [124], Jackson and Watts [62], Konishi and Ray [71], and Dutta, Ghosal, and Ray [38]. We do not touch on the topic of learning in networks, a topic which has been the focus of much work by Goyal [45,46], nor do we discuss random networks, introduced in economics by Kirman [68]. For random networks we refer the reader to the recent book by Vega-Redondo [121] and the references contained therein. Finally, we do not discuss the statistical mechanics of network formation. For this topic, we recommend to the reader the excellent papers by Blume [13] and Durlauf [37].

Because our focus is on foundational issues in strategic network formation, and in particular stability, we do not discuss any of the plethora of economic applications that can be found in the exploding literature on social and economic networks. For now, we can only offer the reader the following modest and incomplete list of topics and papers:

**Development and insurance** Bloch, Genicot, and Ray [12], Bramoulle and Kranton [18];

**Employment and labor markets** Rees [98], Granovetter [50], Boorman [16], Montgomery [86], Topa [118], Calvo-Armengol [19], Calvo-Armengol and Jackson [21], Calvo-Armengol and Jackson [22];

**Industrial organization and R&D** Demange and Henriot [32], Bloch [10], Kranton and Minehart [74], Goyal and Moraga-Gonzalez [49], Goyal and Joshi [47], Belleflamme and Bloch [7], Bloch [11], and Deroian and Gannon [35], Mauleon, Sempere-Monerris, and Vannetelbosch [83], Wang and Watts [123];

**International trade** Casella and Rauch [23,24] Zissimos [2], Goyal and Joshi [48], and Furusawa and Konishi [41];

**Market microstructure** Tesfatsion [116,117], Kirman, Herreiner and Weisbuch [69], Kranton and Mine-

hart [75], Corominas-Bosch [28], and Even-Dar, Kearns, and Suri [40];

**Public goods** Bramoulle and Kranton [17];

**Organizations, coordination, and communication**

Chwe [27], Currarini [29], Demange [34];

**Marketing and advertising** Galeotti and Moraga-Gonzalez [43].

## The Primitives

Our abstract game of network formation rests on four primitives: The feasible set of networks, players' preferences, the rules of network formation, and a dominance relation over feasible networks. In this section, we discuss in detail these four primitives and in the next section, using these primitives we construct our abstract games of network formation.

## Feasible Networks

**Types of Networks** Surprisingly, there is no agreed-upon definition of a network but rather several definitions depending on the application. But what all definitions have in common is a nonempty set of nodes and a precise mathematical description of how nodes are connected. What differentiates these various definitions then are the details of how nodes are connected. We begin with the most elementary notion of a network, the homogeneous linking network, and proceed to a more complex notion, the heterogeneous directed network introduced in [91].

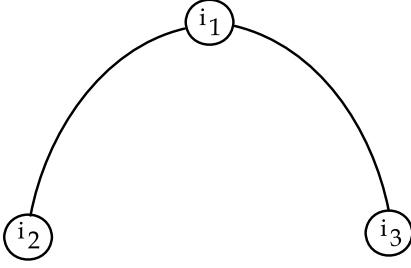
Let  $N$  be a finite set of nodes, with typical element denoted by  $i$ , and let  $A$  be a finite set of link types or arc types, with typical element denoted by  $a$ . If the network is directed (to be defined below), we refer to the elements of  $A$  as arc types, otherwise we will refer to the elements of  $A$  as link types. For any set  $E$ , we denote by  $P(E)$  the collection of all subsets of  $E$ . Finally, for any set  $E$ , we denote by  $|E|$  the cardinality of  $E$  (note that  $|\emptyset| = 0$ ).

### Linking Networks

**Definition 1 (Homogeneous Linking Networks, Myerson [88], Jackson–Wolinsky [64])** Let  $P_2(N)$  denote the set of all subsets of  $N$  of size 2. A linking network,  $G$ , is a subset (possibly empty) of  $P_2(N)$  and for any  $G \subseteq P_2(N)$ , each subset  $\{i, i'\} \in G$  is called a link in  $G$ . The collection of all homogeneous linking networks is denoted by  $P(P_2(N))$ .  $\square$

Thus,  $P_2(N)$  is the set of all possible links and a homogeneous linking network  $G$  is simply a subset of all possible links. For example, if  $N = \{i_1, i_2, i_3\}$ , then

$$P_2(N) = \{\{i_1, i_2\}, \{i_2, i_3\}, \{i_1, i_3\}\},$$



**Networks and Stability, Figure 1**  
**Homogeneous Linking Network  $G_1$**

and the subset

$$G_1 = \{\{i_1, i_2\}, \{i_1, i_3\}\}$$

of  $P_2(N)$  is a homogeneous linking network. Figure 1 depicts homogeneous linking network  $G_1$ .

Here, the link  $\{i_1, i_3\} \in G_1$  denotes that nodes  $i_1$  and  $i_3$  are connected or linked. Note that all links are the same (i. e., links are homogeneous) and links have no orientation or direction. Also, note that in a homogeneous linking network, loops are not allowed by definition (a loop being a link between a node and itself). Finally, note that in a linking network multiple links between any pair of nodes are not allowed. However, because links are homogeneous, multiple links are unnecessary.

The following extended definition of a linking network allows for heterogeneous links. This heterogeneity is represented by a labeling of links using elements of the set  $A$  of link types.

### Definition 2 (Heterogeneous Linking Networks)

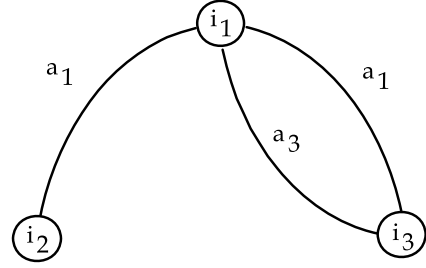
A heterogeneous linking network,  $G$ , is a subset of  $A \times P_2(N)$ . Given any  $G \subseteq A \times P_2(N)$ , each ordered pair  $(a, \{i, i'\}) \in G$  consisting of a link type and a link is called a labeled link in  $G$ . The collection of all heterogeneous linking networks is denoted by  $P(A \times P_2(N))$ .  $\square$

Thus,  $A \times P_2(N)$  is the set of all possible labeled links and a heterogeneous linking network  $G$  is simply a subset of all possible labeled links. For example, given  $N = \{i_1, i_2, i_3\}$  and  $A = \{a_1, a_2, a_3\}$ , the subset

$$G_2 = \{(a_1, \{i_1, i_2\}), (a_1, \{i_1, i_3\}), (a_3, \{i_1, i_3\})\}$$

of  $A \times P_2(N)$  is a heterogeneous linking network. Figure 2 depicts heterogeneous linking network  $G_2$ .

Here, the labeled link  $(a_1, \{i_1, i_3\}) \in G_2$  denotes that nodes  $i_1$  and  $i_3$  are linked by a type  $a_1$  link. First, note that in network  $G_2$  in addition to being linked by an  $a_1$  link, nodes  $i_1$  and  $i_3$  are also linked by an  $a_3$  link. Thus,



**Networks and Stability, Figure 2**  
**Heterogeneous Linking Network  $G_2$**

in a heterogeneous linking network, links are not identical and multiple, distinct links between any given pair of nodes are possible. Second, note that in network  $G_2$ , nodes  $i_1$  and  $i_2$  – like nodes  $i_1$  and  $i_3$  – are linked by an  $a_1$  link. Thus, in a heterogeneous linking network, link types can be used multiple times for different pairs of nodes. Finally, note that in a heterogeneous linking network, links are still without orientation or direction and loops are not possible.

**Directed Networks** The link orientation problem as well as the problem of loops is resolved by moving to directed networks. As is the case with linking networks, there are two categories of directed networks: Homogeneous directed networks and heterogeneous directed networks.

### Definition 3 (Homogeneous Directed Networks)

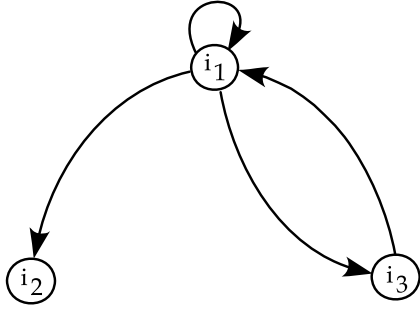
A homogeneous directed network,  $G$ , is a subset of  $N \times N$ . Given any  $G \subseteq N \times N$ , each ordered pair  $(i, i') \in G$  consisting of a beginning node  $i$  and an ending node  $i'$  is called an arc in  $G$ . The collection of all directed networks is denoted by  $P(N \times N)$ .  $\square$

Thus,  $N \times N$  is the set of all possible arcs and a homogeneous directed network  $G$  is simply a subset of all possible arcs. For example, given  $N = \{i_1, i_2, i_3\}$ ,

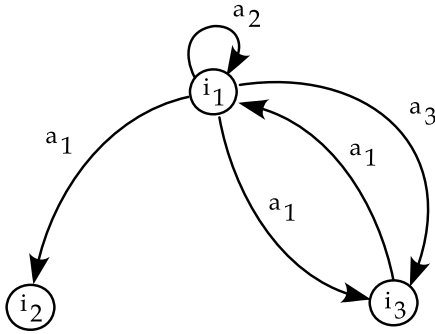
$$G_3 = \{(i_1, i_1), (i_1, i_2), (i_1, i_3), (i_3, i_1)\}$$

is a homogeneous directed network. Figure 3 depicts homogeneous directed network  $G_3$ .

Here, the arc  $(i_1, i_3) \in G_3$  denotes that nodes  $i_1$  and  $i_3$  are connected by an arc *running from node  $i_1$  to node  $i_3$* . Note that because  $(i_3, i_1) \in G_3$  there is also an arc running in the opposite direction from  $i_3$  to  $i_1$ . Also, note that in a homogeneous directed network, while connections have direction, all connections are of the same type – that is, connections are homogeneous. Finally, note that in a directed network loops are allowed. For example,  $(i_1, i_1) \in G_3$  and therefore in network  $G_3$  there is an arc running from node  $i_1$  to node  $i_1$ .



Networks and Stability, Figure 3  
Homogeneous Directed Network  $G_3$



Networks and Stability, Figure 4  
Heterogeneous Directed Network  $G_4$

The following definition, from [94], allows for heterogeneous, multiple arcs by labeling arcs using the set  $A$  of arc types.

**Definition 4 (Heterogeneous Directed Networks)**

A heterogeneous directed network,  $G$ , is a subset of  $A \times (N \times N)$ . Given any  $G \subseteq A \times (N \times N)$ , each ordered pair  $(a, (i, i')) \in G$  consisting of an arc type and an arc is called a labeled arc in  $G$ . The collection of all labeled directed networks is denoted by  $P(A \times (N \times N))$ .  $\square$

Thus,  $A \times (N \times N)$  is the set of all possible labeled arcs and a heterogeneous directed network  $G$  is simply a subset of all possible labeled arcs. For example, given  $N = \{i_1, i_2, i_3\}$  and  $A = \{a_1, a_2, a_3\}$  the subset

$$G_4 = \{(a_2, (i_1, i_1)), (a_1, (i_1, i_2)), (a_1, (i_1, i_3)), (a_1, (i_3, i_1)), (a_3, (i_1, i_3))\}$$

of  $A \times (N \times N)$  is a heterogeneous directed network. Figure 4 depicts heterogeneous directed network  $G_4$ .

Here, the labeled arc  $(a_1, (i_1, i_3)) \in G_4$  denotes that nodes  $i_1$  and  $i_3$  are connected by an arc of type  $a_1$  running from node  $i_1$  to node  $i_3$ . Note that nodes  $i_1$  and  $i_3$  are also connected by an arc of type  $a_3$  running from node  $i_1$

to node  $i_3$ . Thus, in addition to having direction, connections are heterogeneous. Also, note that arc type  $a_1$  is used three times in network  $G_4$ : Once in describing the connection running from  $i_1$  to  $i_2$ , once in describing the connection running from  $i_1$  to  $i_3$ , and once in describing the connection running from  $i_3$  to  $i_1$ . Finally, note that loops are allowed. For example,  $(a_2, (i_1, i_1)) \in G_4$  and therefore in network  $G_4$  there is an  $a_2$  arc running from node  $i_1$  to node  $i_1$ .

*Remarks*

- (1) In the terminology of graph theory (e.g., see Bollobas [15]), a homogeneous linking network is called a graph, while a homogeneous directed network is called a directed graph.
- (2) The following notation is useful in describing heterogeneous directed networks. Given heterogeneous directed network  $G \subseteq A \times (N \times N)$ , let

$$\begin{aligned} G(a) &:= \{(i, i') \in N \times N : (a, (i, i')) \in G\}, \\ G(i, i') &:= \{a \in A : (a, (i, i')) \in G\}, \\ G^+(i) &:= \{a \in A : (a, (i, i')) \in G \text{ for some } i' \in N\}, \\ G^-(i) &:= \{a \in A : (a, (i', i)) \in G \text{ for some } i' \in N\}, \\ G^+(a, i) &:= \{i' \in N : (a, (i, i')) \in G\}, \\ G^-(a, i) &:= \{i' \in N : (a, (i', i)) \in G\}. \end{aligned}$$

For example, referring to heterogeneous directed network  $G_4$  above (see Fig. 4),

$$\begin{aligned} G_4(a_1) &:= \{(i_1, i_2), (i_1, i_3), (i_3, i_1)\}, \\ G_4(i_1, i_3) &:= \{a_1, a_3\}, \\ G_4^+(i_1) &:= \{a_1, a_2, a_3\}, \\ G_4^-(i_1) &:= \{a_1, a_2\}, \\ G_4^+(a_2, i_1) &:= \{i_1\}, \\ G_4^-(a_2, i_1) &:= \{i_1\}. \end{aligned}$$

- (3) The number  $|G^+(i)|$  is the number arc types leaving node  $i$  in network  $G$ , while the number  $|G^+(a, i)|$  is the out degree of node  $i$  for arc types  $a$  in network  $G$ . The number  $|G^-(i)|$  is the number of arc types entering node  $i$  in network  $G$ , while the number  $|G^-(a, i)|$  is the indegree of node  $i$  for arc type  $a$  in network  $G$ . For directed network  $G_4$ , we have for example,

$$\begin{aligned} |G_4^+(i_2)| &= 0 \quad \text{and} \quad |G_4^-(i_2)| = 1, \\ |G_4^+(a_1, i_1)| &= 2 \quad \text{and} \quad |G_4^-(a_1, i_1)| = 1. \end{aligned}$$

- (4) Rockafellar [101], essentially defines a network  $G$  to be a nonempty subset of  $A \times (N \times N)$  such that

for all  $a \in A$ ,  $G(a) \subseteq \{(i, i') \in N \times N: i \neq i'\}$ , and  $|G(a)| \leq 1$ .

**The Feasible Set** In the abstract games of network formation we develop here, we will assume that the game is played over some feasible set of networks  $\mathbb{G}$ . Some examples of feasible sets are: The set of all homogeneous linking networks  $P(P_2(N))$  as in Jackson–Wolinsky [64] and Jackson–van den Nouweland [61]; The set of all homogeneous directed networks  $P(N \times N)$  as in Bala and Goyal [5] and an arbitrary subset of the set of heterogeneous directed networks  $P(A \times (N \times N))$  as in [91] and [93]. The following example is taken from Page and Wooders [92] where the feasible set is taken to be the set of all club networks – a particular class of heterogeneous directed networks.

*Example 1 (Club Networks)* Let  $D$  be a finite set of players with typical element  $d$  and let  $C$  be a finite set of clubs or club locations with typical element  $c$ . As before, let  $A$  be a finite set of arc types. Finally, let  $N = D \cup C$  be the set of nodes. We consider an abstract game of network formation played over the feasible set  $\mathbb{G} \subset P(A \times (N \times N))$  of club networks where  $G \in \mathbb{G}$  if and only if  $G$  is a non-empty subset of  $A \times (D \times C)$  such that (i) for all players  $d \in D$ , the set

$$G(d) := \{(a, c) \in A \times C: (a, (d, c)) \in G\}$$

is nonempty and (ii) for all  $(a, (d, c)) \in G$ ,  $a \in A(d, c)$ . Here,  $A(d, c)$  is the set of actions (represented by arc types) available to player  $d$  in club  $c$ . Given club network  $G \in \mathbb{G}$ ,  $(a, (d, c)) \in G$  means that in club network  $G$  player  $d$  is a member of club  $c$  and takes action  $a \in A(d, c)$  – or in the terminology of directed networks, that in club network  $G$ , there is an arc of type  $a$  running from node (player)  $d$  to node (club)  $c$ . Thus, in this example the feasible set is a set of bipartite directed networks.

We remark that the basic model of club formation underlying this example has a long history in the literature, going back to economies with essentially homogeneous agents modeled as games in characteristic function form (Shubik (bridge game) [110]) and serves as an example of several models in more recent literature on coalitional games (cf., Banerjee, Konishi and Sonmez [6], Bogomolnaia and Jackson [14], Diamantoudi and Xue [36]), and in economies with clubs (cf. Arnold and Wooders [2] and Allouch and Wooders [1]). As in Konishi, Le Breton and Weber [70] and Demange [33], for example, we allow “free entry” into clubs.  $\square$

**Paths and Circuits** A sequence of links  $\{(i, i')_k\}_k$  in  $G \in \mathbb{G} \subseteq P(P_2(N))$  constitutes a path if each link  $\{(i, i')_k\}_k$

has one node in common with the preceding link  $\{(i, i')_{k-1}\}$  and the other node in common with the succeeding link  $\{(i, i')_{k+1}\}$ . A circuit is a finite path  $\{(i, i')_k\}_{k=1}^h$  in  $G$  which begins at a node  $i$  and returns the same node. The length of a path is the number of links in the path.

A sequence of labeled links  $\{(a, \{i, i'\})_k\}_k$  in  $G \in \mathbb{G} \subseteq P(A \times P_2(N))$  constitutes a path if each labeled link  $(a, \{i, i'\})_k$  has one node in common with the preceding labeled link  $(a, \{i, i'\})_{k-1}$  and the other node in common with the succeeding link  $(a, \{i, i'\})_{k+1}$ . A circuit is a finite path  $\{(a, \{i, i'\})_k\}_{k=1}^h$  in  $G$  which begins at a node  $i$  and ends at the same node. The length of a path is the number of labeled links in the path.

A sequence of arcs  $\{(i, i')_k\}_k$  in  $G \in \mathbb{G} \subseteq P(N \times N)$  constitutes a path if the beginning node  $i$  of arc  $(i, i')_k$  coincides with the ending node  $i'$  of preceding arc  $(i, i')_{k-1}$ . A circuit is a finite path  $\{(i, i')_k\}_{k=1}^h$  in  $G$  such that node  $i$  of arc  $(i, i')_1$  and node  $i'$  of arc  $(i, i')_h$  are the same node. The length of a path is the number of arcs in the path.

Finally, a sequence of labeled arcs  $\{(a, (i, i'))_k\}_k$  in  $G \in \mathbb{G} \subseteq P(A \times (N \times N))$  constitutes a path if the beginning node  $i$  of labeled arc  $(a, (i, i'))_k$  coincides with the ending node  $i'$  of preceding arc  $(a, (i, i'))_{k-1}$ . A circuit is a finite path  $\{(a, (i, i'))_k\}_{k=1}^h$  in  $G$  such that node  $i$  of labeled arc  $(a, (i, i'))_1$  and node  $i'$  of labeled arc  $(a, (i, i'))_h$  are the same node. The length of a path is the number of labeled arcs in the path.

In Fig. 5,  $\{(a_1, (i_3, i_1))_1, (a_2, (i_1, i_1))_2, (a_1, (i_1, i_2))_3\}$  is a path in  $G_4$  of length 3, while  $\{(a_1, (i_3, i_1))_1, (a_2, (i_1, i_1))_2, (a_3, (i_1, i_3))_3\}$  is a circuit in  $G_4$  of length 3.

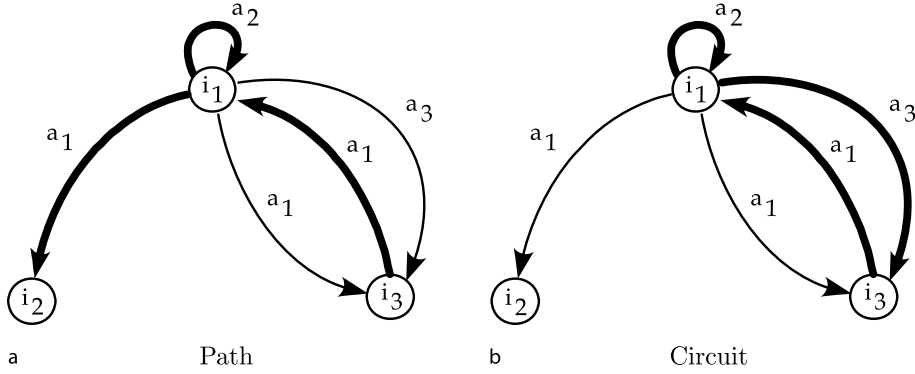
### Players' Preferences

*For the remainder of this entry we will assume that the set of players is given by the set of nodes  $N$ . Thus, henceforth the nodes represent players in the game of network formation.*

Let  $\Gamma(N)$  denote the collection of all coalitions of players (i. e., nonempty subsets of  $N$ ) with typical element denoted by  $S$ .

For each player  $i \in N$  let  $\succ_i$  be an irreflexive binary relation on  $\mathbb{G} (= P(P_2(N)) \text{ or } P(N \times N))$  and write  $G' \succ_i G$  if player  $i \in N$  prefers network  $G' \in \mathbb{G}$  to network  $G \in \mathbb{G}$ . Because  $\succ_i$  is irreflexive,  $G \not\succ_i G$  for all networks  $G \in \mathbb{G}$ . Coalition  $S' \in \Gamma(N)$  prefers network  $G'$  to network  $G$ , written  $G' \succ_{S'} G$ , if  $G' \succ_i G$  for all players  $i \in S'$ . Note that because players' preferences  $\{\succ_i\}_{i \in N}$  are irreflexive, coalitional preferences,  $\{\succ_S\}_{S \in \Gamma(N)}$ , are also irreflexive.

**A Remark on Weak Preferences** Players are said to have weak preferences on  $\mathbb{G} (= P(P_2(N)) \text{ or } P(N \times N))$ , denoted by  $\succeq_i$  if  $G' \succeq_i G$  means that player  $i$  either strongly



Networks and Stability, Figure 5

Path and circuit in Heterogeneous Directed Network  $G_4$ 

prefers  $G'$  to  $G$  (denoted  $G' \succ_i G$ ) or is indifferent between  $G'$  and  $G$  (denoted  $G' \sim_i G$ ). If coalitional preferences are based on weak preference, then we say that coalition  $S' \in P(N)$  *weakly prefers* network  $G'$  to network  $G$ , written  $G' \succ_{wS'} G$ , if for all players  $i \in S'$ ,  $G' \succeq_i G$  and if for at least one player  $i' \in S'$ ,  $G' \succ_{i'} G$ . Note that if preferences are weak and  $G' \succ_{wS'} G$  where  $S'$  consists of a single player  $i'$ , so that  $S' = \{i'\}$ , then  $G' \succ_{i'} G$ . Finally, note that weak coalitional preferences  $\{\succ_{wS}\}_{S \in \Gamma(N)}$  are irreflexive (i. e.,  $G \not\succ_{wS} G$  for all  $G \in \mathbb{G}$  and  $S \in \Gamma(N)$ ).

**Network Payoff Functions** In many applications, players' preferences are specified via real-valued network payoff functions,  $\{v_i(\cdot)\}_{i \in N}$ . If this is the case, then for each player  $i \in N$  and each network  $G \in \mathbb{G}$ ,  $v_i(G)$  is the payoff to player  $i$  in network  $G$ . Note that the payoff  $v_i(G)$  to player  $i$  in network  $G$  depends on the entire network. Thus, the player may be affected by connections between other players even when he himself has no direct or indirect connection with those players. Intuitively, 'widespread' network externalities are allowed.

Given payoff functions  $\{v_i(\cdot)\}_{i \in N}$ , player  $i$  prefers network  $G'$  to network  $G$  if  $v_i(G') > v_i(G)$ . Coalitional preferences can then be specified by stating that coalition  $S' \in \Gamma(N)$  prefers network  $G'$  to network  $G$  if  $v_i(G') > v_i(G)$  for all  $i \in S'$ .

**Preference Supernetworks** By viewing each network  $G$  in feasible set  $\mathbb{G}$  as a node in a larger network, we can represent coalitional preferences as a heterogeneous directed network. To begin, let

$$\mathcal{P} := \{p_S : S \in \Gamma(N)\}$$

denote the set of arc labels for preference arcs (or  $p$ -arcs for short).

**Definition 5 (Coalitional Preference Supernetworks, [94])** Given feasible set  $\mathbb{G}$  ( $= P(P_2(N))$  or  $P(N \times N)$ ), a coalitional preference supernetwork  $\mathbf{P}$  is a subset of  $\mathcal{P} \times (\mathbb{G} \times \mathbb{G})$  such that  $(p_{S'}, (G, G'))$  is contained in  $\mathbf{P}$  if and only if  $G' \succ_{S'} G$ .

### The Rules of Network Formation

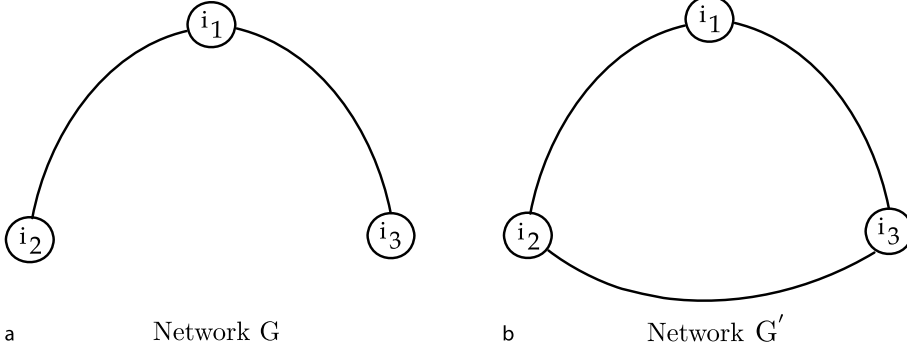
The rules of network formation are specified via a collection of coalitional effectiveness relations  $\{\rightarrow_S\}_{S \in \Gamma(N)}$  defined on the feasible set of networks  $\mathbb{G}$  ( $= P(P_2(N))$  or  $P(N \times N)$ ). Each effectiveness relation  $\rightarrow_S$  represents what a coalition  $S$  can do. Thus, if  $G \rightarrow_S G'$  this means that under the rules of network formation coalition  $S \in \Gamma(N)$  can change network  $G \in \mathbb{G}$  to network  $G' \in \mathbb{G}$  by adding, subtracting, or replacing connections in  $G$  (where, depending on the feasible set, a *connection* is a link or an arc).

### Examples of Network Formation Rules

**Jackson–Wolinsky Rules [64] (Bilateral–Unilateral Rules)** Assume that the feasible set of networks  $\mathbb{G}$  is equal to the set of homogeneous linking networks  $P(P_2(N))$ . Under the Jackson–Wolinsky rules of network formation (see [64]),

- (i) adding a link from player  $i$  to player  $i'$  requires that both players  $i$  and  $i'$  agree to add the link (i. e., link addition is bilateral);
- (ii) subtracting a link from player  $i$  to player  $i'$  requires that player  $i$  or player  $i'$  or both agree to subtract the link (i. e., link subtraction can be unilateral);
- (iii) link addition or link subtraction takes place one link at a time.





**Networks and Stability, Figure 6**  
**a Network G. b Network G'**

Thus, for any pair of networks  $G$  and  $G'$  in  $\mathbb{G}$ , if  $G \rightarrow_S G'$  and  $G \neq G'$ , then

- either  $G' = G \cup \{i, i'\}$  for some  $\{i, i'\} \in P_2(N)$   
 and  $S = \{i, i'\}$   
 or  $G' = G \setminus \{i, i'\}$  for some  $\{i, i'\} \in G$   
 and  $S = \{i\}$  or  $S = \{i'\}$  or  $S = \{i, i'\}$ .

To illustrate, consider Fig. 6 depicting two homogeneous linking networks  $G$  and  $G'$ .

Observe that

$$G' = G \cup \{i_2, i_3\} \quad \text{and} \quad G = G' \setminus \{i_2, i_3\}.$$

Under the effectiveness relations implied by the Jackson–Wolinsky rules, for networks  $G$  and  $G'$  we have

$$G \xrightarrow[\{i_2, i_3\}]{} G', \quad G' \xrightarrow[\{i_2, i_3\}]{} G, \quad G' \xrightarrow[\{i_2\}]{} G, \quad G' \xrightarrow[\{i_3\}]{} G.$$

**Jackson–van den Nouweland Rules [61] (Bilateral–Unilateral Rules)** Again assume that the feasible set of networks  $\mathbb{G}$  is equal to the set of homogeneous linking networks  $P(P_2(N))$ . Under the Jackson–van den Nouweland rules of network formation,

- (i) adding a link from player  $i$  to player  $i'$  requires that both players  $i$  and  $i'$  agree to add the link (i.e., link addition is bilateral);
- (ii) subtracting a link from player  $i$  to player  $i'$  requires that player  $i$  or player  $i'$  or both agree to subtract the link (i.e., link subtraction can be unilateral).

Thus, the Jackson–van den Nouweland rules are the Jackson–Wolinsky rules without the one-link-at-a-time restriction. Note that if link addition is bilateral and link subtraction is unilateral (i.e., if rules (i) and (ii) hold), then

$G \rightarrow_S G''$  and  $G \neq G''$  implies that

- (i) if  $\{i, i'\} \in G''$  and  $\{i, i'\} \notin G$ , then  $\{i, i'\} \subseteq S$ ;

and

- (ii) if  $\{i, i'\} \notin G''$  and  $\{i, i'\} \in G$ ,  
 then  $\{i, i'\} \cap S \neq \emptyset$ .

To illustrate, consider Fig. 7 depicting two homogeneous linking networks  $G$  and  $G''$ .

Observe that

$$G'' = (G \setminus \{i_1, i_3\}) \cup \{i_2, i_3\}$$

and

$$G = (G'' \setminus \{i_2, i_3\}) \cup \{i_1, i_3\}.$$

Under the effectiveness relations implied by the Jackson–van den Nouweland rules, for networks  $G$  and  $G''$  we have

$$G \xrightarrow[\{i_1, i_2, i_3\}]{} G'', \quad G \xrightarrow[\{i_2, i_3\}]{} G'', \quad G'' \xrightarrow[\{i_1, i_3\}]{} G, \quad G'' \xrightarrow[\{i_1, i_2, i_3\}]{} G.$$

Note that under the one-link-at-a-time restriction, it is not possible under the Jackson–Wolinsky rules to move directly from network  $G$  to network  $G''$  or directly from network  $G''$  to network  $G$  (i.e.,  $G$  and  $G''$  are *not* related under the effectiveness relations  $\{\rightarrow_S\}_{S \in \Gamma(N)}$ ). Instead, under the Jackson–Wolinsky rules, the change from  $G$  to  $G''$  or from  $G''$  to  $G$  requires two moves. For example,

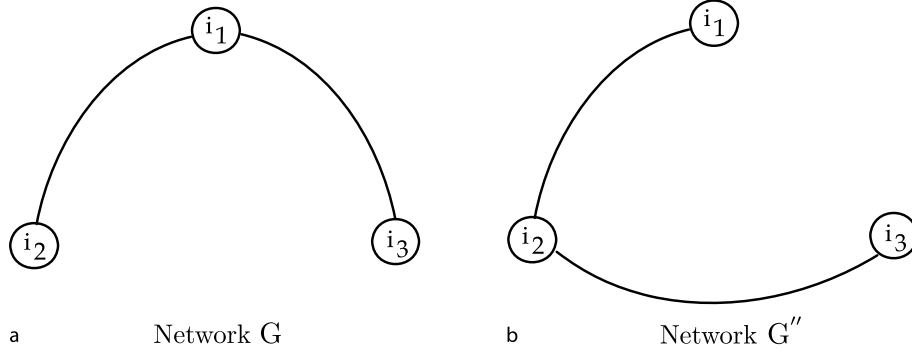
$$\text{first } G \xrightarrow[\{i_2, i_3\}]{} G', \quad \text{and then,}$$

$$G' \xrightarrow[\{i_3\}]{} G'' \quad \text{or} \quad G' \xrightarrow[\{i_1\}]{} G'';$$

or

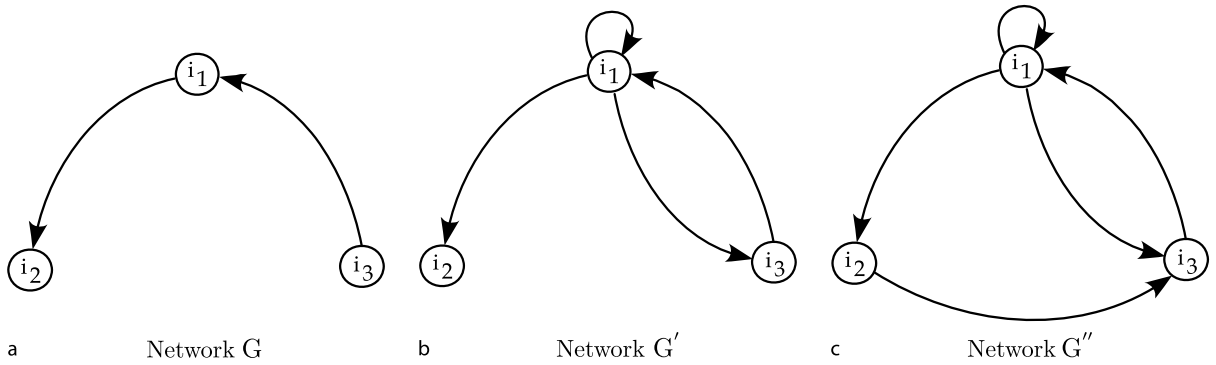
$$\text{first } G'' \xrightarrow[\{i_1, i_3\}]{} G', \quad \text{and then,}$$

$$G' \xrightarrow[\{i_3\}]{} G \quad \text{or} \quad G' \xrightarrow[\{i_2\}]{} G.$$



Networks and Stability, Figure 7

a Network G. b Network G''



Networks and Stability, Figure 8

a Network G. b Network G'. c Network G''

**Bala–Goyal Rules** [5] (*Noncooperative Rules – Unilateral–Unilateral Rules*) Now assume that the feasible set of networks is equal to the set of homogeneous directed networks  $P(N \times N)$ . Translating Bala and Goyal rules into our notation and terminology,

- (i) adding an arc from player  $i$  to player  $i'$  requires only that player  $i$  agree to add the arc (i. e., arc addition is unilateral and can be carried out only by the initiator, player  $i$ );
- (ii) subtracting an arc from player  $i$  to player  $i'$  requires only that player  $i$  agree to subtract the arc (i. e., arc subtraction is unilateral and can be carried out only by the initiator, player  $i$ );
- (iii)  $G \rightarrow_S G'$  implies that  $|S| = 1$  (i. e., only network changes brought about by individual players are allowed).

We shall also refer to rules (i)–(iii) as noncooperative. Note that a player  $i$  can add or subtract an arc to player  $i'$  without regard to the preferences of player  $i'$  and can add and/or subtract arcs to several players simultaneously

and can do so without regard to those players' preferences. Thus in general under noncooperative rules, effectiveness relations display a type of symmetry, and in particular, if  $G \xrightarrow[\{i\}]{} G'$ , then  $G' \xrightarrow[\{i\}]{} G$ .

To illustrate, consider Fig. 8 depicting three homogeneous directed networks  $G$ ,  $G'$ , and  $G''$ .

Under the effectiveness relations implied by noncooperative rules for networks  $G$  and  $G'$  in Fig. 8 we have

$$G \xrightarrow[\{i_1\}]{} G', \quad G' \xrightarrow[\{i_1\}]{} G.$$

Note that under noncooperative rules, networks  $G$  and  $G''$  in Fig. 8 are *not* related under the effectiveness relations  $\{\rightarrow_{\{i\}}\}_{i \in N}$ . However, under the noncooperative rules we have, for example, the following effectiveness relations

$$G \rightarrow_{\{i_1\}} G', \quad G' \rightarrow_{\{i_2\}} G''$$

and

$$G'' \rightarrow_{\{i_2\}} G', \quad G' \rightarrow_{\{i_1\}} G.$$

**Rules Supernetworks** Again by viewing each network  $G$  in feasible set  $\mathbb{G}$  as a node in a larger network, we can represent the rules of network formation as a heterogeneous directed network. To begin, let

$$\mathcal{M} := \{m_S : S \in \Gamma(N)\}$$

denote the set of arc labels for move arcs (or  $m$ -arcs for short).

**Definition 6 (Rules Supernetworks, [94])** Given feasible set  $\mathbb{G} (= P(P_2(N))$  or  $P(N \times N)$ ), a rules supernetwork  $\mathbf{R}_\rho$  is a subset of  $\mathcal{M} \times (\mathbb{G} \times \mathbb{G})$  such that  $(m_{S'}, (G, G'))$  is contained in  $\mathbf{R}_\rho$  if and only if  $G \rightarrow_{S'} G'$ , where  $\rho$  denotes the name of the network formation rules in force. We shall adopt the convention that  $\rho = jw$  if the rules are Jackson–Wolinsky,  $\rho = jn$  if the rules are Jackson–van den Nouweland, and  $\rho = bg$  if the rules are Bala–Goyal or noncooperative.  $\square$

### Supernetworks

Given feasible set  $\mathbb{G} (= P(P_2(N))$  or  $P(N \times N)$ ), coalitional preferences  $\{\succ_S\}_{S \in \Gamma(N)}$ , and coalitional effectiveness relations  $\{\rightarrow_S\}_{S \in \Gamma(N)}$  can be represented by a heterogeneous directed network called a supernetwork (see [89] and [94]). In particular, given preference supernetwork  $\mathbf{P}$  and rules supernetwork  $\mathbf{R}_\rho$ , the corresponding supernetwork is given by

$$\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho.$$

Letting  $\mathcal{A} := \mathcal{P} \cup \mathcal{M}$  (i. e., the union of all preference arcs and move arcs), then

$$\mathbf{G}_\rho \subset \mathcal{A} \times (\mathbb{G} \times \mathbb{G}).$$

### Dominance Relations

**Direct Dominance** Given feasible set  $\mathbb{G} (= P(P_2(N))$  or  $P(N \times N)$ ), coalitional preferences  $\{\succ_S\}_{S \in \Gamma(N)}$ , and coalitional effectiveness relations  $\{\rightarrow_S\}_{S \in \Gamma(N)}$ , network  $G' \in \mathbb{G}$  *directly dominates* network  $G \in \mathbb{G}$ , written  $G' \triangleright G$ , if for some coalition  $S' \in \Gamma(N)$ ,

$$G \prec_{S'} G' \quad \text{and} \quad G \xrightarrow{S'} G'.$$

Thus, network  $G'$  directly dominates network  $G$  if some coalition  $S'$  prefers  $G'$  to  $G$  and if under the rules of network formation coalition  $S'$  has the power to change  $G$  to  $G'$ .

Note that direct dominance is irreflexive but not in general transitive. Also note that if  $\mathbf{G}_\rho$  is the supernetwork, then  $G' \triangleright G$  if and only if  $(p_{S'}, (G, G')) \in \mathbf{G}_\rho$  and  $(m_{S'}, (G, G')) \in \mathbf{G}_\rho$ , for some coalition  $S'$ .

**Indirect Dominance** Given feasible set  $\mathbb{G} (= P(P_2(N))$  or  $P(N \times N)$ ), coalitional preferences  $\{\succ_S\}_{S \in \Gamma(N)}$ , and coalitional effectiveness relations  $\{\rightarrow_S\}_{S \in \Gamma(N)}$ , network  $G' \in \mathbb{G}$  *indirectly dominates* network  $G \in \mathbb{G}$ , written  $G' \triangleright\triangleright G$ , if there is a *finite* sequence of networks,

$$G_0, G_1, \dots, G_h,$$

with  $G = G_0$ ,  $G' = G_h$ , and  $G_k \in \mathbb{G}$  for  $k = 0, 1, \dots, h$ , and a corresponding sequence of coalitions,

$$S_1, S_2, \dots, S_h,$$

such that for  $k = 1, 2, \dots, h$

$$G_{k-1} \xrightarrow{S_k} G_k, \quad \text{and} \quad G_{k-1} \prec_{S_k} G_h.$$

Note that if network  $G'$  indirectly dominates network  $G$  (i. e., if  $G' \triangleright\triangleright G$ ), then what matters to the initially deviating coalition  $S_1$ , as well as all the coalitions along the way, is that the ultimate network outcome  $G' = G_h$  be preferred. Thus, for example, the initially deviating coalition  $S_1$  will not be deterred from changing network  $G_0$  to network  $G_1$  even if network  $G_1$  is not preferred to network  $G = G_0$ , as long as the ultimate network outcome  $G' = G_h$  is preferred to  $G_0$ , that is, as long as  $G_0 \prec_{S_1} G_h$ . Finally, note that indirect dominance is irreflexive but not in general transitive.

In order to capture the idea of farsightedness in strategic behavior, Chwe [26] analyzed abstract games equipped with indirect dominance relations in great detail, introducing the equilibrium notions of consistency and largest consistent set. The basic idea of indirect dominance goes back to the work of Guilbaud [53] and Harsanyi [54].

Given the supernetwork representation of preferences and rules,  $\mathbf{G}_\rho$ , we can write,  $G' \triangleright\triangleright G$  if there is a *finite* sequence of networks,

$$G_0, G_1, \dots, G_h,$$

with  $G = G_0$ ,  $G' = G_h$ , and  $G_k \in \mathbb{G}$  for  $k = 0, 1, \dots, h$ , and a corresponding sequence of coalitions,

$$S_1, S_2, \dots, S_h,$$

such that for  $k = 1, 2, \dots, h$

$$(m_{S_k}, (G_{k-1}, G_k)) \in \mathbf{G}_\rho,$$

and

$$(p_{S_k}, (G_{k-1}, G_h)) \in \mathbf{G}_\rho.$$

**Path Dominance** Any irreflexive dominance relation  $>$  on  $\mathbb{G}$  ( $= P(P_2(N))$  or  $P(N \times N)$ ) – for example, direct or indirect – induces a path dominance relation on the set of networks (sometimes referred to as the transitive closure of  $>$ ). In particular, corresponding to dominance relation  $>$  on networks  $\mathbb{G}$  there is a corresponding path dominance relation  $\geq_p$  on  $\mathbb{G}$  specified as follows: Network  $G' \in \mathbb{G}$  path dominates network  $G \in \mathbb{G}$  with respect to  $>$  (i.e., with respect to the underlying dominance relation  $>$ ), written  $G' \geq_p G$ , if  $G' = G$  or if there exists a *finite* sequence of networks  $\{G_k\}_{k=0}^h$  in  $\mathbb{G}$  with  $G_h = G'$  and  $G_0 = G$  such that for  $k = 1, 2, \dots, h$

$$G_k > G_{k-1}.$$

We refer to such a finite sequence of networks as a *finite domination path* and we say network  $G'$  is *>-reachable* from network  $G$  if there exists a finite domination path from  $G$  to  $G'$ . Thus,

$$G' \geq_p G \quad \text{if and only if} \quad \begin{cases} G' \text{ is } >\text{-reachable from} \\ G, \text{ or} \\ G' = G. \end{cases}$$

Note that, even though the underlying dominance relation  $>$  is irreflexive and intransitive or transitive, the induced path dominance relation  $\geq_p$  on  $\mathbb{G}$  is both reflexive ( $G \geq_p G$ ) and transitive ( $G' \geq_p G$  and  $G'' \geq_p G'$  implies that  $G'' \geq_p G$ ).

**>-Supernetworks** Let  $>$  denote the irreflexive dominance relation on  $\mathbb{G}$ . It is often useful to represent  $>$  as a *homogeneous* directed network,  $\mathbf{D}_>$ , where  $\mathbf{D}_>$  is a subset of  $\mathbb{G} \times \mathbb{G}$  and where  $>\text{-arc } (G, G') \in \mathbf{D}_>$  if and only if  $G' > G$  (i.e., if and only if  $G' >\text{-dominates } G$ ). We call such a homogeneous directed network (or directed graph) a  $>\text{-supernetwork}$ . For example, suppose  $\mathbb{G} = \{G_1, G_2, G_3, \dots, G_7\}$  and suppose the dominance relation  $>$  on  $\mathbb{G}$  is a direct dominance relation and has the supernetwork representation given in Fig. 9.

Note that network  $G_5$  is  $>\text{-reachable}$  through  $\mathbf{D}_>$  from network  $G_1$  by the domination path given by the  $>\text{-arc}$  sequence

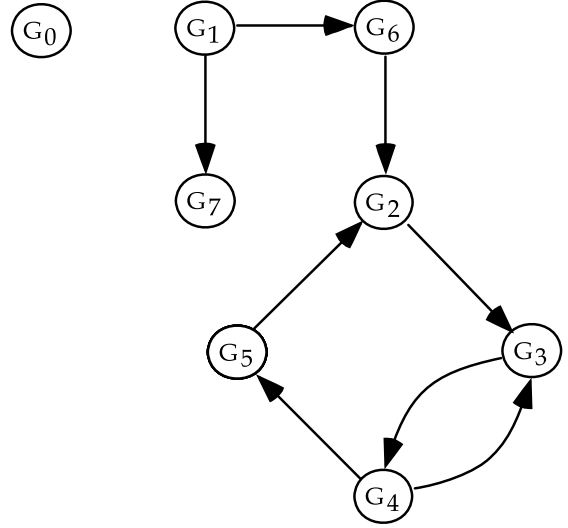
$$\{(G_1, G_6)_1, (G_6, G_2)_2, (G_2, G_3)_3, (G_3, G_4)_4, (G_4, G_5)_5\}.$$

Thus,  $G_5$  path dominates  $G_1$ . Note that network  $G_2$  is  $>\text{-reachable}$  from network  $G_2$  by the domination circuit given by the  $>\text{-arc}$  sequence

$$\{(G_2, G_3)_1, (G_3, G_4)_2, (G_4, G_5)_3, (G_5, G_2)_4\}.$$

and that network  $G_3$  is  $>\text{-reachable}$  from network  $G_3$  by two domination circuits given by the  $>\text{-arc}$  sequences

$$\{(G_3, G_4)_1, (G_4, G_5)_2, (G_5, G_2)_3, (G_2, G_3)_4\}$$



**Networks and Stability, Figure 9**  
 $>\text{-supernetwork } \mathbf{D}_>$

and

$$\{(G_3, G_4)_1, (G_4, G_3)_2\}.$$

Because networks  $G_2$  and  $G_5$  are on the same circuit,  $G_5$  is  $>\text{-reachable}$  from  $G_2$  and  $G_2$  is  $>\text{-reachable}$  from  $G_5$ . Thus,  $G_5$  path dominates  $G_2$  (i.e.,  $G_5 \geq_p G_2$ ) and  $G_2$  path dominates  $G_5$  (i.e.,  $G_2 \geq_p G_5$ ). The same cannot be said of networks  $G_1$  and  $G_5$ . In particular, while  $G_5 \geq_p G_1$ , it is not true that  $G_1 \geq_p G_5$  because  $G_1$  is not  $>\text{-reachable}$  from  $G_5$ . Finally, note that network  $G_0$  is isolated in  $\mathbf{D}_>$ . In particular,  $G_0$  is *not* reachable through  $\mathbf{D}_>$  from any network in  $\mathbb{G}$  and no network in  $\mathbb{G}$  is reachable through  $\mathbf{D}_>$  from  $G_0$ . In general, a network  $G \in \mathbb{G}$  is isolated if there does not exist a network  $G' \in \mathbb{G}$  with  $G' \geq_p G$  or  $G \geq_p G'$ .

Note that if the direct dominance relation with  $>\text{-supernetwork}$  depicted in Fig. 9 has underlying coalitional preferences  $\{\succ_S\}_{S \in \Gamma(N)}$  and coalitional effectiveness relations  $\{\rightarrow_S\}_{S \in \Gamma(N)}$ , then the  $>\text{-arc}$  from network  $G_3$  to network  $G_4$  in Fig. 9 means that for some coalition  $S$ ,  $G_4$  is preferred to  $G_3$  and more importantly, that coalition  $S$  has the power to change network  $G_3$  to network  $G_4$ . Thus,  $G_3 \prec_{sS} G_4$  and  $G_3 \rightarrow_S G_4$ . But because there is a  $>\text{-arc}$  in the opposite direction, from network  $G_4$  to network  $G_3$ ,  $G_3$  also directly dominates  $G_4$ . Thus for some coalition  $S'$  disjoint from coalition  $S$  ( $S' \cap S = \emptyset$ ),  $G_4 \prec_{s'S'} G_3$  and  $G_4 \rightarrow_{S'} G_3$ . Finally, note that if coalitional preferences over networks are weak (i.e., are based on weak preferences), then the statement, 'for some coalition  $S'$  disjoint from coalition  $S$  can be weakened to for some coalition  $S'$  not equal to coalition  $S$ . With this weakening, the require-

ment that the intersection of  $S$  and  $S'$  be empty is no longer needed.

### Abstract Games of Network Formation and Stability

An abstract game of network formation consists of a feasible set of networks equipped with a dominance relation. We shall consider two classes of games: (i) games where the feasible set of networks  $\mathbb{G}$  ( $= P(P_2(N))$  or  $P(N \times N)$ ) is equipped with an irreflexive dominance relation  $>$ , either direct or indirect (i. e.,  $>$  is equal to  $\triangleright$  or  $\triangleright\triangleright$ ), induced by coalitional preferences and network formation rules; and (ii) games where the feasible set of networks is equipped with a path dominance relation  $\geq_p$  induced by such an irreflexive dominance relation.

### Network Formation Games with Respect to Irreflexive Dominance

In this section we consider the abstract game with respect to irreflexive dominance given by the pair

$$(\mathbb{G}, >).$$

Throughout this section we will assume that primitives are represented by supernetwork  $\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho$  (where  $\rho$  is equal to  $iw$ ,  $jn$ , or  $bg$ ) and  $>$ -supernetwork  $\mathbf{D}_>$  (where  $>$  is equal to  $\triangleright$  or  $\triangleright\triangleright$ ), and that the feasible set of networks  $\mathbb{G}$  is equal to the set of homogeneous linking networks  $P(P_2(N))$  or the set of homogeneous directed networks  $P(N \times N)$ .

**Quasi-Stability and Stability** We define the  $>$ -distance from  $G_0$  to  $G_1$  in  $\mathbf{D}_>$  to be the length of the *shortest*  $>$ -path from  $G_0$  to  $G_1$  if  $G_1$  is  $>$ -reachable from  $G_0$  in  $\mathbf{D}_>$ , and  $+\infty$  if  $G_1$  is not reachable from  $G_0$  in  $\mathbf{D}_>$ . We denote the distance from  $G_0$  to  $G_1$  in  $\mathbf{D}_>$  by  $d_{\mathbf{D}_>}(G_0, G_1)$ . Thus,

$$d_{\mathbf{D}_>}(G_0, G_1) := \begin{cases} \text{length of shortest } >\text{-path} \\ \text{from } G_0 \text{ to } G_1 \text{ in } \mathbf{D}_>, & \text{if } G_1 \text{ is reachable from } G_0, \\ +\infty, & \text{if } G_1 \text{ is not reachable from } G_0 \text{ in } \mathbf{D}_>. \end{cases}$$

The following are network renditions of quasi-stable and stable sets.

### Definition 7 (Quasi-Stable Sets and Stable Sets, [8,25])

- (1) A subset  $\mathbb{Q}$  of networks in  $\mathbb{G}$  is said to be quasi-stable for network formation game  $(\mathbb{G}, >)$  if,
  - (a)  $\mathbb{Q}$  is internally stable, that is,  $d_{\mathbf{D}_>}(G_0, G_1) \geq 2$ , whenever  $G_0$  and  $G_1$  are in  $\mathbb{Q}$ , with  $G_0 \neq G_1$ , and

- (b)  $\mathbb{Q}$  is externally quasi-stable, that is, given any  $G_0 \notin \mathbb{Q}$ , there exists  $G_1 \in \mathbb{Q}$  with  $d_{\mathbf{D}_>}(G_0, G_1) \leq 2$ .

- (2) A subset  $\mathbb{S}$  of networks in  $\mathbb{G}$  is said to be stable for network formation game  $(\mathbb{G}, >)$  if,

- (a)  $\mathbb{S}$  is internally stable and
- (b)  $\mathbb{S}$  is externally stable, that is, given any  $G_0 \notin \mathbb{S}$ , there exists  $G_1 \in \mathbb{S}$  with  $d_{\mathbf{D}_>}(G_0, G_1) \leq 1$ .  $\square$

Thus, if  $\mathbb{Q}$  is externally quasi-stable, a path of length at most 2 is required to get from any network outside of  $\mathbb{Q}$  to a network in  $\mathbb{Q}$ , whereas, if  $\mathbb{Q}$  is externally stable a path of length at most 1 is required.

Letting

$$P_{>}(G_0) := \{G \in \mathbb{G} : G > G_0\},$$

an alternative way to write part (2) of the definition above is as follows:

- (2)' A subset  $\mathbb{S}$  of networks in  $\mathbb{G}$  is said to be stable if

- (a) (internal stability)  $G \in \mathbb{S}$  implies that  $P_{>}(G) \cap \mathbb{S} = \emptyset$  and
- (b) (external stability)  $G \notin \mathbb{S}$  implies that  $P_{>}(G) \cap \mathbb{S} \neq \emptyset$ .

If  $\mathbb{S}$  is stable (or quasi-stable), then it is automatically non-empty. Note that a stable set  $\mathbb{S}$  is simply a von Neuman–Morgenstern stable set with respect to the dominance relation  $>$  defined on  $\mathbb{G}$ . Also, note that if  $\mathbb{S}$  is stable then it is automatically quasi-stable.

We now state a remarkably simple result on the existence of quasi-stable sets. This result is a network rendition of a general result due to Chvatal and Lovasz [25] on the existence of quasi-stable sets in directed graphs (here the directed graph is the  $>$ -supernetwork  $\mathbf{D}_>$ ; also see, Galeana-Sanchez and Li [42]).

**Theorem 1 (Existence of quasi-stable sets for network formation games, [89])** *There exists a quasi-stable set  $\mathbb{Q}$  for network formation game  $(\mathbb{G}, >)$ .*

In fact, it follows from the Theorem due to Chvatal and Lovasz [25] that any finite set  $\mathbb{Z}$  equipped with an irreflexive binary relation  $<$  has a  $<$ -quasi-stable set. Moreover, if the relation  $<$  is transitive, then any  $<$ -quasi-stable set is  $<$ -stable.

Next we state two results on the existence of stable sets.

**Theorem 2 (Existence of stable sets for network formation games, [89])**

- (1) If  $\mathbf{D}_>$  contains no  $>$ -circuits, then there exists a unique stable  $\mathbb{S}$  for  $(\mathbb{G}, >)$ .
- (2) If  $\mathbf{D}_>$  contains no  $>$ -circuits of odd length, then there exists a stable  $\mathbb{S}$  for  $(\mathbb{G}, >)$ .



Part (1) of Theorem 2 is an immediate consequence of a 1958 result due to Berge (see Theorem 4, p. 48 in Berge [8]). Part (2) of Theorem 2 is a supernetwork version of the classical result due to Richardson [100]. A  $>$ -circuit in supernetwork  $\mathbf{D}_>$  is said to be of odd length if there is an odd number of connections in the circuit.

**Farsighted Consistency** Chwe [26] in an influential paper introduced the notion of farsighted consistency in an abstract game. The following is a definition of farsighted consistency for abstract games of network formation.

**Definition 8 (Farsighted Consistency, [89])** A subset  $\mathbb{F}$  of networks in  $\mathbb{G}$  is said to be farsightedly consistent for network formation game  $(\mathbb{G}, \triangleright \triangleright)$  if,

$$\begin{aligned} & \text{for all } G_0 \in \mathbb{F}, \\ & (m_{S_1}, (G_0, G_1)) \in \mathbf{G}_\rho \text{ for some } G_1 \in \mathbb{G} \\ & \text{and some coalition } S_1, \text{ implies that} \\ & \text{there exists } G_2 \in \mathbb{F} \\ & \text{with } G_2 = G_1 \text{ or } G_2 \triangleright \triangleright G_1 \text{ such that,} \\ & (p_{S_1}, (G_0, G_2)) \notin \mathbf{G}_\rho. \end{aligned}$$

□

In words, a subset of directed networks  $\mathbb{F}$  is said to be farsightedly consistent if given any network  $G_0 \in \mathbb{F}$  and any  $m_{S_1}$ -deviation to network  $G_1 \in \mathbb{G}$  by coalition  $S_1$  (via adding, subtracting, or replacing arcs in accordance with  $\mathbf{R}_\rho$ ) there exists further deviations leading to some network  $G_2 \in \mathbb{F}$  where the initially deviating coalition  $S_1$  is not better off – and possibly worse off. A network  $G \in \mathbb{G}$  is said to be farsightedly consistent if  $G \in \mathbb{F}$  where  $\mathbb{F}$  is a farsightedly consistent set.

If (i)  $\mathbb{G} = P(P_2(N))$ , (ii) coalitional preferences are weak, denoted by  $\{>_{wS}\}_{S \in \Gamma(N)}$ , so that indirect dominance is weak (denoted by  $\triangleright \triangleright_w$  and defined in the obvious way), and (iii) coalitional effectiveness relations are determined by Jackson–Wolinsky rules, then the notion of farsighted consistency above (essentially due to Chwe [26]) is closely related to the notion of pairwise farsighted stability introduced in Herrings, Mauleon, and Vannetelbosch [55].

For any game  $(\mathbb{G}, \triangleright \triangleright)$ , there can be many farsightedly consistent sets. We shall denote by  $\mathbb{F}^*$  the *largest farsightedly consistent set*. Thus, if  $\mathbb{F}$  is farsightedly consistent, then  $\mathbb{F} \subseteq \mathbb{F}^*$ . Unlike quasi-stable sets and stable sets where existence implies nonemptiness, in considering farsightedly consistent sets two critical questions arise: (i) does there exist a largest farsightedly consistent set of networks for  $(\mathbb{G}, \triangleright \triangleright)$ , and (ii) is it nonempty? Our next result provides a positive answer to both questions.

**Theorem 3 (Existence, Uniqueness, and Nonemptiness of  $\mathbb{F}^*$ , [94])** *There exists a unique, nonempty largest farsightedly consistent set  $\mathbb{F}^*$  for network formation game  $(\mathbb{G}, \triangleright \triangleright)$ . Moreover,  $\mathbb{F}^*$  is externally stable; that is, if network  $G$  is not contained in  $\mathbb{F}^*$ , then there exists a network  $G'$  contained in  $\mathbb{F}^*$  that indirectly dominates  $G$  (i.e.,  $G' \triangleright \triangleright G$ ).*

The method of proving existence and uniqueness is a straightforward, supernetwork rendition of Chwe's [26] method and is similar to the method introduced by Roth [103,104]. Page and Kamat [89] provide an alternative proof (to that of Chwe and of [94]) of the nonemptiness and external stability of the largest consistent set (with respect to indirect dominance). In particular, Page and Kamat modify the indirect dominance relation so as to make it transitive as well as irreflexive. They then show that the unique stable set with respect to path dominance induced by this new transitive indirect dominance relation is contained in the largest farsightedly consistent set – and in this way show that the largest farsightedly consistent set is nonempty and externally stable.

### Network Formation Games with Respect to Path Dominance

In this section we consider the network formation game with respect to path dominance given by the pair

$$(\mathbb{G}, \geq_\rho).$$

Throughout this section we will assume that the underlying primitives,

$$(\mathbb{G}, \{>_S\}, \{\rightarrow_S\}, >)_{S \in \Gamma(N)},$$

are such that  $\mathbb{G}$  is equal to the set of homogeneous linking networks  $P(P_2(N))$  or the set of homogeneous directed networks  $P(N \times N)$ , that the dominance relation  $>$  on  $\mathbb{G}$  is given by either direct dominance  $\triangleright$  or indirect dominance  $\triangleright \triangleright$ , and that primitives are represented by supernetwork  $\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho$  (with  $\rho$  equal to  $jw$ ,  $jn$ , or  $bg$ ) and  $>$ -supernetwork  $\mathbf{D}_>$ .

We will present three notions of stability introduced in [91] and [93] for abstract games of network formation with respect to path dominance over heterogeneous directed networks: (i) strategic basins of attraction, (ii) path dominance stable sets, and (iii) the path dominance core.

### Preliminaries

*Networks Without Descendants* If  $G_1 \geq_\rho G_0$  and  $G_0 \geq_\rho G_1$ , networks  $G_1$  and  $G_0$  are *equivalent*, written

$G_1 \equiv_p G_0$ . If networks  $G_1$  and  $G_0$  are equivalent then either networks  $G_1$  and  $G_0$  coincide or  $G_1$  and  $G_0$  are on the same circuit (see Fig. 9 above for a picture of a circuit). If  $G_1 \geq_p G_0$  but  $G_1$  and  $G_0$  are not equivalent (i. e., not  $G_1 \equiv_p G_0$ ), then network  $G_1$  is a *descendant* of network  $G_0$  and we write

$$G_1 >_p G_0 .$$

Referring to Fig. 9, observe that network  $G_5$  is a descendant of network  $G_1$ , that is,  $G_5 >_p G_1$ .

Network  $G' \in \mathbb{G}$  has no descendants in  $\mathbb{G}$  if for any network  $G \in \mathbb{G}$

$$G \geq_p G' \text{ implies that } G \equiv_p G' .$$

Thus, if  $G'$  has no descendants then  $G \geq_p G'$  implies that  $G$  and  $G'$  coincide or lie on the same circuit. Note that any isolated network is by definition a network without descendants (e. g., network  $G_0$  in Fig. 9).

In attempting to identify and characterize stable homogeneous networks, networks *without descendants* are of particular interest. Here is our main result concerning networks without descendants.

**Theorem 4 (All path dominance network formation games have networks without descendants, [91,93])** *In network formation game  $(\mathbb{G}, \geq_p)$  every network  $G \in \mathbb{G}$  is path dominated by a network  $G' \in \mathbb{G}$  without descendants (i. e.,  $G' \geq_p G$  and  $G'$  has no descendants).*

By Theorem 4, in network formation game  $(\mathbb{G}, \geq_p)$ , corresponding to any network  $G \in \mathbb{G}$  there is a network  $G' \in \mathbb{G}$  without descendants which is  $>$ -reachable from  $G$ . Thus, in any network formation game the set of networks without descendants is nonempty. Referring to Fig. 9, the set of networks without descendants is given by

$$\{G_0, G_2, G_3, G_4, G_5, G_7\} .$$

We shall denote by  $\mathbb{Z}$  the set of networks without descendants.

**Basins of Attraction** Stated loosely, a basin of attraction is a set of *equivalent* networks to which the strategic network formation process represented by the game might tend and from which there is no escape. Formally, we have the following definition.

**Definition 9 (Basin of Attraction, [91,93])** A set of networks  $\mathbb{A} \subseteq \mathbb{G}$  is said to be a basin of attraction for  $(\mathbb{G}, \geq_p)$  if

- (a) the networks contained in  $\mathbb{A}$  are equivalent (i. e., for all  $G'$  and  $G$  in  $\mathbb{A}$ ,  $G' \equiv_p G$ ) and for no set  $\mathbb{A}'$  having

$\mathbb{A}$  as a strict subset is this true that all the networks in  $\mathbb{A}'$  are equivalent, and

- (b) no network in  $\mathbb{A}$  has descendants (i. e., there does not exist a network  $G' \in \mathbb{G}$  such that  $G' >_p G$  for some  $G \in \mathbb{A}$ ).

$\mathbb{A}$  is a strict subset of  $\mathbb{A}'$  if  $\mathbb{A} \subset \mathbb{A}'$  and  $\mathbb{A}' \setminus \mathbb{A} \neq \emptyset$ .

As the following characterization result shows, there is a very close connection between networks without descendants and basins of attraction.

**Theorem 5 (A characterization of basins of attraction, [91,93])** *Let  $\mathbb{A}$  be a subset of networks in  $\mathbb{G}$ . The following statements are equivalent:*

- (1)  $\mathbb{A}$  is a basin of attraction for  $(\mathbb{G}, \geq_p)$ .
- (2) There exists a network without descendants,  $G \in \mathbb{Z}$ , such that

$$\mathbb{A} = \{G' \in \mathbb{Z} : G' \equiv_p G\} .$$

In light of Theorem 5, we conclude that in any network formation game  $(\mathbb{G}, \geq_p)$ ,  $\mathbb{G}$  contains a *unique*, finite, disjoint collection of basins of attraction, say  $\{\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_m\}$ , where for each  $k = 1, 2, \dots, m$  ( $m \geq 1$ )

$$\mathbb{A}_k = \mathbb{A}_G := \{G' \in \mathbb{Z} : G' \equiv_p G\}$$

for some network  $G \in \mathbb{Z}$ . Note that for networks  $G'$  and  $G$  in  $\mathbb{Z}$  such that  $G' \equiv_p G$ ,  $\mathbb{A}_{G'} = \mathbb{A}_G$  (i. e. the basins of attraction  $\mathbb{A}_{G'}$  and  $\mathbb{A}_G$  coincide). Also, note that if network  $G \in \mathbb{G}$  is isolated, then  $G \in \mathbb{Z}$  and

$$\mathbb{A}_G := \{G' \in \mathbb{Z} : G' \equiv_p G\} = \{G\}$$

is, by definition, a basin of attraction – but a very uninteresting one.

**Example 2 (Basins of Attraction)** In Fig. 9 above the set of networks without descendants is given by

$$\mathbb{Z} = \{G_0, G_2, G_3, G_4, G_5, G_7\} .$$

Even though there are six networks without descendants, because networks  $G_2, G_3, G_4$ , and  $G_5$  are equivalent, there are only three basins of attraction:

$$\begin{aligned} \mathbb{A}_1 &= \{G_0\} , & \mathbb{A}_2 &= \{G_2, G_3, G_4, G_5\} , \\ & & \mathbb{A}_3 &= \{G_7\} . \end{aligned}$$

Moreover, because  $G_2, G_3, G_4$ , and  $G_5$  are equivalent,

$$\mathbb{A}_{G_2} = \mathbb{A}_{G_3} = \mathbb{A}_{G_4} = \mathbb{A}_{G_5} = \{G_2, G_3, G_4, G_5\} .$$

□

**Stable Sets with Respect to Path Dominance** The formal definition of a  $\geq_p$ -stable set is as follows.

**Definition 10 (Stable Sets with Respect to Path Dominance, [91,93])** A subset  $\mathbb{V}$  of networks in  $\mathbb{G}$  is said to be a stable set for  $(\mathbb{G}, \geq_p)$  if

- (a) (internal  $\geq_p$ -stability) whenever  $G_0$  and  $G_1$  are in  $\mathbb{V}$ , with  $G_0 \neq G_1$ , then neither  $G_1 \geq_p G_0$  nor  $G_0 \geq_p G_1$  hold, and
- (b) (external  $\geq_p$ -stability) for any  $G_0 \notin \mathbb{V}$  there exists  $G_1 \in \mathbb{V}$  such that  $G_1 \geq_p G_0$ .  $\square$

In other words, a nonempty subset of networks  $\mathbb{V}$  is a stable set for  $(\mathbb{G}, \geq_p)$  if  $G_0$  and  $G_1$  are in  $\mathbb{V}$ , with  $G_0 \neq G_1$ , then  $G_1$  is not  $>$ -reachable from  $G_0$ , nor is  $G_0 >$ -reachable from  $G_1$ , and if  $G_0 \notin \mathbb{V}$ , then there exists  $G_1 \in \mathbb{V}$  reachable from  $G_0$ .

We now have our main results on the existence, construction, and cardinality of stable sets. These results can be viewed as variations on some classical results from graph theory applied to network formation games (e.g., see Berge [8], Chap. 2).

**Theorem 6 (Stable sets: Existence, construction, and cardinality, [91,93])** Without loss of generality assume that  $(\mathbb{G}, \geq_p)$  has basins of attraction given by

$$\{\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_m\},$$

where basin of attraction  $\mathbb{A}_k$  contains  $|\mathbb{A}_k|$  many networks (i.e.,  $|\mathbb{A}_k|$  is the cardinality of  $\mathbb{A}_k$ ). Then the following statements are true:

- (1)  $\mathbb{V} \subseteq \mathbb{G}$  is a stable set for  $(\mathbb{G}, \geq_p)$  if and only if  $\mathbb{V}$  is constructed by choosing one network from each basin of attraction, that is, if and only if  $\mathbb{V}$  is of the form

$$\mathbb{V} = \{G_1, G_2, \dots, G_m\},$$

where  $G_k \in \mathbb{A}_k$  for  $k = 1, 2, \dots, m$ .

- (2)  $(\mathbb{G}, \geq_p)$  possesses

$$|\mathbb{A}_1| \cdot |\mathbb{A}_2| \cdot \dots \cdot |\mathbb{A}_m| := M$$

many stable sets and each stable set,  $\mathbb{V}_q$ ,  $q = 1, 2, \dots, M$ , has cardinality

$$|\mathbb{V}_q| = |\{\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_m\}| = m.$$

**Example 3 (Basins of Attraction and Stable Sets)** Referring to Fig. 9, it follows from Theorem 6 that because

$$|\mathbb{A}_1| \cdot |\mathbb{A}_2| \cdot |\mathbb{A}_3| = 1 \cdot 4 \cdot 1 = 4,$$

the network formation game  $(\mathbb{G}, \geq_p)$  has 4 stable sets, each with cardinality 3. By examining Fig. 9 in light of Theorem 6, we see that the stable sets for  $(\mathbb{G}, \geq_p)$  are given by

$$\begin{aligned} \mathbb{V}_1 &= \{G_0, G_2, G_7\}, & \mathbb{V}_2 &= \{G_0, G_3, G_7\}, \\ \mathbb{V}_3 &= \{G_0, G_4, G_7\}, & \mathbb{V}_4 &= \{G_0, G_5, G_7\}. \end{aligned}$$

$\square$

It should be noted that by equipping the abstract network formation game with the path dominance relation rather than the original dominance relation, we avoid the famous Lucas [78] example of a game with no stable set.

### The Path Dominance Core

**Definition 11 (The Path Dominance Core, [91,93])**

A network  $G \in \mathbb{G}$  is contained in the path dominance core  $\mathbb{C}$  of network formation game  $(\mathbb{G}, \geq_p)$  if and only if there does not exist a network  $G' \in \mathbb{G}$ ,  $G' \neq G$ , such that  $G' \geq_p G$ .

Our next results give necessary and sufficient conditions for the path dominance core of a network formation game over homogeneous networks to be nonempty, as well as a recipe for constructing the path dominance core.

**Theorem 7 (Path dominance core: Nonemptiness and construction, [91,93])** Without loss of generality assume that  $(\mathbb{G}, \geq_p)$  has basins of attraction given by

$$\{\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_m\},$$

where basin of attraction  $\mathbb{A}_k$  contains  $|\mathbb{A}_k|$  many networks. Then the following statements are true:

- (1)  $(\mathbb{G}, \geq_p)$  has a nonempty path dominance core if and only if there exists a basin of attraction containing a single network, that is, if and only if for some basin of attraction  $\mathbb{A}_k$ ,  $|\mathbb{A}_k| = 1$ .
- (2) Let

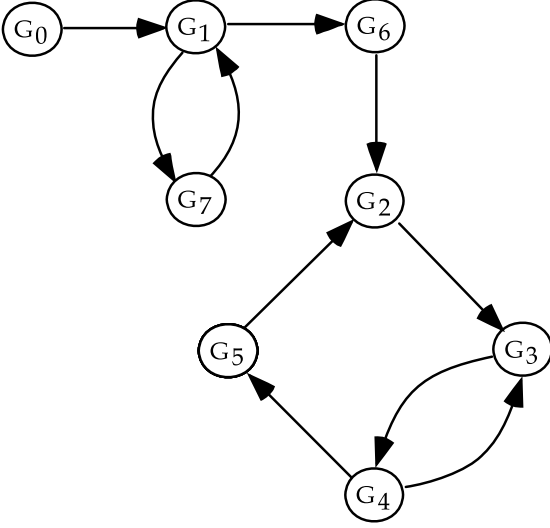
$$\{\mathbb{A}_{k_1}, \mathbb{A}_{k_2}, \dots, \mathbb{A}_{k_n}\} \subseteq \{\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_m\},$$

be the subset of basins of attraction containing all basins having cardinality 1. Then the path dominance core  $\mathbb{C}$  of  $(\mathbb{G}, \geq_p)$  is given by

$$\mathbb{C} = \{G_{k_1}, G_{k_2}, \dots, G_{k_n}\},$$

where  $G_{k_i} \in \mathbb{A}_{k_i}$ , for  $i = 1, 2, \dots, n$ .

If coalitional preferences over networks are based on weak preferences, that is, if coalitional preferences are given by  $\{\succ_{wS}\}_{S \in \Gamma(N)}$ , then the corresponding path dominance



Networks and Stability, Figure 10  
A different  $\succ$ -supernetwork  $D_{\succ}$

core – the weak path dominance core – is contained in the path dominance core based on strong preference relations.

*Example 4 (Basins of Attraction and the Path Dominance Core, [91,93])* It follows from Theorem 7 that the path dominance core of the network formation game  $(\mathbb{G}, \geq_p)$  with feasible set  $\mathbb{G} = \{G_0, G_1, \dots, G_7\}$  and path dominance relation  $\geq_p$  induced by the dominance relation depicted in Fig. 9 is

$$\mathbb{C} = \{G_0, G_7\}.$$

Suppose that the  $\succ$ -supernetwork corresponding to the direct dominance relation  $\succ$  on  $\mathbb{G} = \{G_0, G_1, \dots, G_7\}$  is instead depicted by Fig. 10.

Now the network formation game  $(\mathbb{G}, \geq_p)$  has 3 circuits and 1 basin of attraction,  $\mathbb{A} = \{G_2, G_3, G_4, G_5\}$ . Because  $|\mathbb{A}_1| = 4$ , by Theorem 7 the path dominance core of  $(\mathbb{G}, \geq_p)$  is empty. By Theorem 6,  $(\mathbb{G}, \geq_p)$  has 4 stable sets each containing 1 network. These stable sets are given by

$$\mathbb{V}_1 = \{G_2\}, \mathbb{V}_2 = \{G_3\}, \mathbb{V}_3 = \{G_4\}, \mathbb{V}_4 = \{G_5\}.$$

□

**The Path Dominance Core and Constrained Pareto Efficiency** We say that a network  $G \in \mathbb{G}$  is constrained Pareto efficient for game  $(\mathbb{G}, \geq_p)$  if and only if there does not exist another network  $G' \in \mathbb{G}$  such that (i) some coalition  $S$  can change network  $G$  to network  $G'$  (that is,  $G \rightarrow_S G'$  for some coalition  $S \in \Gamma(N)$ ) and (ii)  $G'$  is preferred by all players (that is,  $G <_i G'$  for all players  $i \in N$ ).

Letting  $\mathbb{E}$  denote the set of all constrained Pareto efficient networks, it is easy to see that the path dominance core  $\mathbb{C}$  of network formation game  $(\mathbb{G}, \geq_p)$  is a subset of  $\mathbb{E}$ , that is,  $\mathbb{C} \subseteq \mathbb{E}$ .

Under the classical notion of Pareto efficiency, a network  $G$  is said to be Pareto efficient if and only if there does not exist another network  $G'$  such that  $G <_i G'$  for all players  $i \in N$ , regardless of whether or not some coalition  $S$  can change network  $G$  to network  $G'$ . Letting  $\mathbb{PE}$  denote the set of all classically Pareto efficient networks, it is easy to see that  $\mathbb{PE} \subseteq \mathbb{E}$ . Note, however, that if under the rules of network formation, any network  $G$  can be changed to any other network  $G'$  via the actions of some coalition  $S$ , then the notions of constrained Pareto efficiency and classical Pareto efficiency are equivalent. Thus, if the collection of coalitional effectiveness relations  $\{\rightarrow_S\}_{S \in \Gamma(N)}$  on  $\mathbb{G}$  is complete, that is, if for any pair of networks  $G$  and  $G'$  in  $\mathbb{G}$ ,  $G \rightarrow_S G'$  for some coalition  $S \in \Gamma(N)$ , then  $\mathbb{PE} = \mathbb{E}$ , and we have  $\mathbb{C} \subseteq \mathbb{PE} = \mathbb{E}$ .

### Strong Stability, Pairwise Stability, Nash Stability, and Farsighted Consistency

In this section we continue our discussion of the network formation game with respect to path dominance given by the pair

$$(\mathbb{G}, \geq_p).$$

Throughout this section we will continue to assume that the underlying primitives,

$$(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, \succ)_{S \in \Gamma(N)},$$

are such that  $\mathbb{G}$  is equal to the set of homogeneous linking networks  $P(P_2(N))$  or the set of homogeneous directed networks  $P(N \times N)$ , that the dominance relation  $\succ$  on  $\mathbb{G}$  is given by either direct dominance  $\triangleright$  or indirect dominance  $\triangleright\triangleright$ , and that primitives are represented by supernetwork  $\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho$  (with  $\rho$  equal to  $jw$ ,  $jn$ , or  $bg$ ) and  $\succ$ -supernetwork  $\mathbf{D}_{\succ}$ .

We will complete our unification of stability results for network formation games by concluding that, depending on how we further specialize the primitives underlying the game  $(\mathbb{G}, \geq_p)$ , the path dominance core is equal to the set of pairwise stable networks [64], the set of strongly stable networks [39,61], or the set of Nash networks [5]. We also present results on the relationships between basins of attraction, the path dominance core, and the largest farsightedly consistent set [26]. All of these results follow immediately from more general results in Page and Wooders [91,93] where the notions of strong stability, pairwise

stability, Nash stability, and farsighted consistency are all extended to heterogeneous directed networks.

### Strongly Stable Homogeneous Networks

We begin with a formal definition of strong stability based on that of Jackson–van den Nouweland [61].

**Definition 12 (Strong Stability, [91,93])** Assume that  $\mathbb{G}$  is equal to the set of all homogeneous linking networks,  $P(P_2(N))$ , and that the Jackson–van den Nouweland rules are in force. Network  $G \in \mathbb{G}$  is said to be strongly stable in  $(\mathbb{G}, \geq_p)$  if for all  $G' \in \mathbb{G}$  and  $S \in \Gamma(N)$ ,  $G \rightarrow_S G'$  implies that  $G \not\prec_S G'$ .  $\square$

Thus, a network is strongly stable if whenever a coalition has the power to change the network to another network, the coalition will be deterred from doing so because the change is not preferred by the coalition. If coalitional preferences are strong, the change ‘not being preferred’ means that the change will not make *all* members of the coalition better off. If coalitional preferences are weak (i. e., based on weak preferences), the change ‘not being preferred’ means that the change will either make *no* members better off or will make some members better off and some members worse off. Note that under our definition of strong stability a network  $G \in \mathbb{G}$  that cannot be changed to another network by any coalition is strongly stable.

If (i) coalitional preferences are weak (i. e.,  $\{\succ_{wS}\}_{S \in \Gamma(N)}$ ), and (ii) coalitional effectiveness relations are determined by Jackson–van den Nouweland rules, then the definition of strong stability above is exactly that of Jackson–van den Nouweland. As it stands, our definition is closely related to that given by Dutta–Mutuswami [39].

We now have our main result on the path dominance core and strong stability. Denote the set of strongly stable networks by  $\mathbb{SS}$ .

**Theorem 8 (The path dominance core and strong stability, [91,93])** Assume that  $\mathbb{G}$  is equal to the set of all homogeneous linking networks,  $P(P_2(N))$ , and that the Jackson–van den Nouweland rules are in force.

- (1) If the path dominance core  $\mathbb{C}$  of  $(\mathbb{G}, \geq_p)$  is nonempty, then  $\mathbb{SS}$  is nonempty and  $\mathbb{C} \subseteq \mathbb{SS}$ .
- (2) If the dominance relation  $>$  underlying  $\geq_p$  is a direct dominance relation  $\triangleright$ , then  $\mathbb{C} = \mathbb{SS}$  and  $\mathbb{SS}$  is nonempty if and only if there exists a basin of attraction containing a single network.

Note that the set of strongly stable homogeneous linking networks is contained in the set of constrained

Pareto efficient homogeneous linking networks. Thus,  $\mathbb{C} \subseteq \mathbb{SS} \subseteq \mathbb{E}$ .

### Pairwise Stable Networks

The following definition of pairwise stability is a translation of the Jackson–Wolinsky definition [64].

**Definition 13 (Pairwise Stability, [91,93])** Assume that  $\mathbb{G}$  is equal to the set of all homogeneous linking networks,  $P(P_2(N))$ , and that the Jackson–Wolinsky rules are in force. Network  $G \in \mathbb{G}$  is said to be pairwise stable in  $(\mathbb{G}, \geq_p)$  if for all  $\{i, i'\} \in P_2(N)$ ,

- (1)  $G \rightarrow_{\{i, i'\}} G \cup \{i, i'\}$  implies that  $G \not\prec_{\{i, i'\}} G \cup \{i, i'\}$ ;
- (2)(a)  $G \rightarrow_{\{i\}} G \setminus \{i, i'\}$  implies that  $G \not\prec_{\{i\}} G \setminus \{i, i'\}$ , and
- (2)(b)  $G \rightarrow_{\{i'\}} G \setminus \{i, i'\}$  implies that  $G \not\prec_{\{i'\}} G \setminus \{i, i'\}$ .

$\square$

Thus, a homogeneous linking network is pairwise stable if there is no incentive for any pair of players to add a link to the existing network and there is no incentive for any player who is party to a link in the existing network to dissolve or remove the link. Note that under our definition of pairwise stability a network  $G \in \mathbb{G}$  that cannot be changed to another network by any coalition, or can only be changed by coalitions of size greater than 2, is pairwise stable.

Let  $\mathbb{PS}$  denote the set of pairwise stable networks. It follows from the definitions of strong stability and pairwise stability that  $\mathbb{SS} \subseteq \mathbb{PS}$ . Moreover, if the full set of Jackson–Wolinsky rules are in force, then  $\mathbb{SS} = \mathbb{PS}$ . Jackson–van den Nouweland [61] provide two examples of the potential for strong stability to refine pairwise stability (i. e., two examples where  $\mathbb{SS}$  is a strict subset of  $\mathbb{PS}$ ). However, under Jackson–Wolinsky rules because network changes can occur only one link at a time and because deviations by coalitions of more than two players are not possible such refinements are not possible driving  $\mathbb{SS}$  and  $\mathbb{PS}$  to equality.

We now have our main result on the path dominance core and pairwise stability.

**Theorem 9 (The path dominance core and pairwise stability, [91,93])** Assume that  $\mathbb{G}$  is equal to the set of all homogeneous linking networks,  $P(P_2(N))$ , and that the Jackson–Wolinsky rules are in force.

- (1) If the path dominance core  $\mathbb{C}$  of  $(\mathbb{G}, \geq_p)$  is nonempty, then  $\mathbb{PS}$  is nonempty and  $\mathbb{C} \subseteq \mathbb{PS}$ .



(2) If the dominance relation  $>$  underlying  $\geq_p$  is a direct dominance relation  $\triangleright$ , then  $\mathbb{C} = \mathbb{PS}$  and  $\mathbb{PS}$  is non-empty if and only if there exists a basin of attraction containing a single network.

Theorem 9 can be viewed as an extension of a result due to Jackson and Watts [62] on the existence of pairwise stable homogeneous linking networks for network formation games induced by Jackson–Wolinsky rules. In particular, Jackson and Watts [62] show that for this particular class of Jackson–Wolinsky network formation games, if there does not exist a closed cycle of networks, then there exists a pairwise stable network. Our notion of a strategic basin of attraction containing *multiple* networks corresponds to their notion of a closed cycle of networks. Thus, stated in our terminology, Jackson and Watts show that for this class of network formation games, if there does not exist a basin of attraction containing multiple networks, then there exists a pairwise stable network. Following our approach, by part 2 of Theorem 9 the existence of *at least one* strategic basin containing a single network is both *necessary and sufficient* for the existence of a pairwise stable network.

### Nash Networks

The following definition of Nash networks is a variation on the definition from Bala and Goyal [5].

**Definition 14 (Nash Networks, [91,93])** Assume that  $\mathbb{G}$  is equal to the set of all homogeneous directed networks,  $P(N \times N)$ , and that the Bala–Goyal rules are in force. Network  $G \in \mathbb{G}$  is said to be a Nash network in  $(\mathbb{G}, \geq_p)$  if for all  $G' \in \mathbb{G}$ ,  $G \rightarrow_{\{i\}} G'$  implies that  $G \not\prec_{\{i\}} G'$ .  $\square$

Thus, a homogeneous directed network is Nash if whenever an individual player has the power to change the network to another network, the player will have no incentive to do so. We shall denote by  $\mathbb{NE}$  the set of Nash networks. Note that under our definition any network that cannot be changed to another network by a coalition of size 1 is a Nash network. Finally, note that the set of strongly stable networks  $\mathbb{SS}$  is contained in the set of Nash networks  $\mathbb{NE}$ .

We now have our main result on the path dominance core and Nash stability.

**Theorem 10 (The path dominance core and Nash networks, [91,93])** Assume that  $\mathbb{G}$  is equal to the set of all homogeneous directed networks,  $P(N \times N)$ , and that the Bala–Goyal rules are in force.

(1) If the path dominance core  $\mathbb{C}$  of  $(\mathbb{G}, \geq_p)$  is nonempty, then  $\mathbb{NE}$  is nonempty and  $\mathbb{C} \subseteq \mathbb{NE}$ .

(2) If the dominance relation  $>$  underlying  $\geq_p$  is a direct dominance relation  $\triangleright$ , then  $\mathbb{C} = \mathbb{NE}$  and  $\mathbb{NE}$  is non-empty if and only if there exists a basin of attraction containing a single network.

We close this section by noting that under the Bala–Goyal rules the set of Nash networks  $\mathbb{NE}$  is contained in the set of constrained Pareto efficient networks  $\mathbb{E}$ . If in addition the dominance relation is direct, then  $\mathbb{C} = \mathbb{SS} = \mathbb{NE} \subseteq \mathbb{E}$ .

### Farsightedly Consistent Networks

Our final result summarizes the relationships between basins of attraction, the path dominance core, and the largest farsightedly consistent set.

**Theorem 11 (Basins of attraction, the path dominance core, and the largest consistent set, [91,93])** Assume that (i)  $\mathbb{G}$  is equal to the set of homogeneous linking networks  $P(P_2(N))$  and that the Jackson–Wolinsky rules or the Jackson–van den Nouweland rules are in force; or (ii) that  $\mathbb{G}$  is equal to the set of homogeneous directed networks  $P(N \times N)$  and that the Bala–Goyal rules are in force.

Given network formation game  $(\mathbb{G}, \geq_p)$ , where path dominance is induced by an indirect dominance relation  $\triangleright \triangleright$ , assume without loss of generality that  $(\mathbb{G}, \geq_p)$  has nonempty largest consistent set given by  $\mathbb{F}^*$  and basins of attraction given by

$$\{\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_m\}.$$

Then the following statements are true:

(1) Each basin of attraction  $\mathbb{A}_k$ ,  $k = 1, 2, \dots, m$ , has a nonempty intersection with the largest consistent set  $\mathbb{F}^*$ , that is

$$\mathbb{F}^* \cap \mathbb{A}_k \neq \emptyset, \quad \text{for } k = 1, 2, \dots, m.$$

(2) If  $(\mathbb{G}, \geq_p)$  has a nonempty path dominance core  $\mathbb{C}$ , then

$$\mathbb{C} \subseteq \mathbb{F}^*.$$

### Singleton Basins of Attraction

In the abstract games,  $(\mathbb{G}, \geq_p)$ , that we have considered, the key condition guaranteeing nonemptiness of the path dominance core is the existence of basins of attraction containing a single network. Question: Are there classes of games for which this is true? In general, if the irreflexive dominance relation  $>$  inducing path dominance  $\geq_p$

is transitive, then the  $>$ -supernetwork  $\mathbf{D}_>$  is without circuits, and therefore all basins of attraction for the game  $(\mathbb{G}, \geq_p)$  contain a single network. Unfortunately, if the dominance relation is given by direct or indirect dominance, then transitivity fails to hold in general. In the next two sections we identify several classes of network formation games having singleton basins.

### Network Formation Games and Potential Functions

Assume that  $\mathbb{G}$  is equal to the set of homogeneous directed networks,  $P(N \times N)$ , and that the Bala–Goyal rules are in force, so that primitives are represented by supernetwork  $\mathbf{G}_{bg} := \mathbf{P} \cup \mathbf{R}_{bg}$ . In addition assume that player preferences over  $P(N \times N)$  are specified via payoff functions  $\{v_i(\cdot)\}_{i \in N}$  and that the dominance relation  $>$  over  $P(N \times N)$  is given by direct dominance  $\triangleright$ . Thus,  $G' \triangleright G$  if and only if for some player  $i' \in N$ ,  $G \xrightarrow{\{i'\}} G'$  and  $v_{i'}(G') > v_{i'}(G)$ .

We say that the noncooperative network formation game  $(\mathbb{G}, \geq_p)$  is a potential game if there exists a function

$$P(\cdot): \mathbb{G} \rightarrow R$$

such that for all  $G$  and  $G'$  with  $G \xrightarrow{\{i'\}} G'$  for some player  $i'$ ,

$$v_{i'}(G') > v_{i'}(G) \quad \text{if and only if} \quad P(G') > P(G).$$

It is easy to see that any noncooperative network formation game  $(\mathbb{G}, \geq_p)$  possessing a potential function (i. e., a potential game) has no circuits, and thus possesses strategic basins of attraction each consisting of a single network. Thus, we can conclude from our Theorem 8 that any noncooperative network formation game possessing a potential function has a nonempty path dominance core. In addition, we know from our Theorem 10 that in this example the path dominance core  $\mathbb{C}$  is equal to the set of Nash networks  $\mathbb{NE}$ .

As has been shown by Monderer and Shapley [85], potential games are closely related to congestion games introduced by Rosenthal [102]. Page and Wooders [92] introduce a club network formation game which is a variant of the noncooperative network formation game described above – but for a class of heterogeneous directed networks – and using methods similar to those introduced by Hollar [57], show that this game possesses a potential function. Prior papers studying potential games in the context of linking networks include Qin [97], Slikker, Dutta, van den Nouweland, and Tijs [112] and Slikker and van den Nouweland [114]. These papers have focused on providing

the strategic underpinnings of the Myerson value (Myerson [88] and Aumann and Myerson [4]).

### Jackson–Wolinsky Network Formation Games

Assume that  $\mathbb{G}$  is equal to the set of all homogeneous linking networks,  $P(P_2(N))$ , and that the Jackson–van den Nouweland rules are in force, so that rules are represented by rules supernetwork  $\mathbf{R}_{jn}$ . In addition assume that player preferences over  $P(P_2(N))$  are weak and therefore that coalitional preferences,  $\{>_{wS}\}_{S \in \Gamma(N)}$ , are weak. Finally, assume that the dominance relation  $>$  on  $P(P_2(N))$  is given by direct dominance – but because coalitional preferences are weak, direct dominance is weak, denoted by  $\triangleright_w$ .

In the Jackson–Wolinsky network formation game coalitional preferences are specified by player payoff functions,  $\{v_i(\cdot)\}_{i \in N}$ , and player payoff functions are in turn specified by a network value function

$$v(\cdot): \mathbb{G} \rightarrow R$$

together with an allocation rule,  $Y(G, v) = (Y_i(G, v))_{i \in N} \in R^{|N|}$  satisfying

$$\sum_{i \in N} Y_i(G, v) = v(G).$$

Thus in the Jackson–Wolinsky game, each player's payoff function  $v_i(\cdot)$  is given by  $Y_i(\cdot, v)$  where  $v(\cdot)$  is the network value function. The basic idea here is that given network  $G$ ,  $v(G)$  is the total value generated by network  $G$  and  $Y_i(G, v)$  is value allocated to player  $i$ . Translating Jackson–Wolinsky into our abstract game model, if  $G' \triangleright_w G$  then one of the following is true:

- (i)  $G \xrightarrow{\{i, i'\}} G'$  where  $G' = G \cup \{i, i'\}$  (a link between players  $i$  and  $i'$  is added) and  $Y_i(G', v) \geq Y_i(G, v)$  and  $Y_{i'}(G', v) \geq Y_{i'}(G, v)$  with strict inequality for at least one of the players;
- (ii)  $G \xrightarrow{\{i\}} G'$  where  $G' = G \setminus \{i, i'\}$  (a link between players  $i$  and  $i'$  is subtracted) and  $Y_i(G', v) > Y_i(G, v)$ ;
- (iii)  $G \xrightarrow{\{i'\}} G'$  where  $G' = G \setminus \{i, i'\}$  (a link between players  $i$  and  $i'$  is subtracted) and  $Y_{i'}(G', v) > Y_{i'}(G, v)$ .

Each homogeneous linking network  $G$  can be partitioned into a collection of subnetworks called components as follows. Let  $H \subseteq G$  be any subnetwork of  $G$  and define

$$N_H := \{i \in N: \exists i' \in N \text{ such that } \{i, i'\} \in H\}.$$

Subnetwork  $H \subseteq G$  is said to be a component if

- $i$  and  $i'$  are in  $N_H$ , then there is a path between  $i$  and  $i'$ ; and
- if  $i \in N_H$ ,  $H \subseteq G$ , and  $\{i, i'\} \in G$  then  $\{i, i'\} \in H$ .

Let  $C(G)$  denote the set of all components of  $G$ .

If the allocation rule  $(Y_i(G, v))_{i \in N}$  is *egalitarian* (see [64]); that is, if

$$v_i(G) = Y_i(G, v) = \frac{v(G)}{|N|},$$

then any network  $G^* \in \arg \max_{G \in \mathbb{G}} v(G)$  is pairwise stable and hence by part (2) of Theorem 9,  $(\mathbb{G}, \geq_p)$  has a basin of attraction containing a single network.

Alternatively, if the allocation rule  $(Y_i(G, v))_{i \in N}$  is *componentwise egalitarian* (see [64]); that is, if

$$v_i(G) = Y_i(G, v) = \frac{v(H)}{|N_H|}, \quad \text{for } H \in C(G) \quad \text{and } i \in N_H,$$

then there exists a pairwise stable network (see Jackson [59]), and hence by part (2) of Theorem 9,  $(\mathbb{G}, \geq_p)$  has a basin of attraction containing a single network.

Finally, Jackson [59] has shown that if the allocation rule is given by the Myerson value (see [88] and [4]); that is if

$$v_i(G) = Y_i(G, v) = \sum_{S \subset N \setminus \{i\}} \left( v_{G_{S \cup \{i\}}} - v_{G_S} \left( \frac{|S|!(|N| - |S| - 1)!}{|N|!} \right) \right),$$

where

$$G_S := \{\{i, i'\} \in G : i \in S \text{ and } i' \in S\},$$

then  $(\mathbb{G}, \geq_p)$  has a basin of attraction containing a single network, and hence by part 2 of Theorem 9  $(\mathbb{G}, \geq_p)$  has at least one pairwise stable network.

## Future Directions

There are many possible directions for future research on the topic of networks and stability. Here we only mention a few that, from the perspective of this entry, seem especially promising. There are a number of potential questions to be addressed concerning the path dominance core with direct or indirect dominance. For example, what is the relationship, if any, between basins of attraction and the path dominance core and partnered (or separating) collections of coalitions, as in for example Maschler and

Peleg [81], Maschler, Peleg and Shapley [82], Reny and Wooders [99] and Page and Wooders [90]? Or what is relationship between basins of attraction and the path dominance core and the inner core, as in for example Qin [95,96]?

One of the most pressing issues in our view is strategic network dynamics. Future research will address the following open question: Given the rules of network formation, the preferences of individuals, the strategic behavior of coalitions, and the trembles of nature, what network dynamics are likely to emerge and persist?

Another direction is large networks, where there are many, but still a finite number of nodes or networks with a continuum of nodes. As in the framework of cooperative games (cf., Kovalenkov and Wooders [72] and Wooders, [125,127] for cores of transferable utility and non-transferable utility games) does it hold that some notion of the approximate path dominance core is nonempty if the numbers of players is sufficiently large? Do networks tend towards having some property analogous to the equal treatment property (as in, for example, Debreu and Scarf [31] and Green [52] for exchange economies, [72] or Wooders [126], for cooperative games or Gravel and Thoron [51] for local public goods economies, or Jackson and Watts [63] for a repeated game approach to a matching model). Then there is the problem of characterizing strategic behavior in large networks (as in, for example, Kalai [65] or Wooders, Cartwright and Selten [128]). Under what conditions and to what extent might Kalai's "ex-post stability" or Wooders, Cartwright and Selten's social conformity continue to hold in strategic network formation?

## Acknowledgments

This paper was begun while Page and Wooders were visiting CERMSEM at the University of Paris 1 in June and October of 2007. The authors thank CERMSEM and Paris 1 for their hospitality. URLs: <http://mypage.iu.edu/~fpage>, <http://www.myrnawooders.com>.

## Bibliography

1. Allouch N, Wooders M (2007) Price taking equilibrium in economies with multiple memberships in clubs and unbounded club sizes. J Econ Theory. doi:10.1016/j.jet.2007.07.06
2. Arnold T, Wooders M (2006) Club formation with coordination. University of Warwick Working Paper 640
3. Aumann RJ (1964) Markets with a continuum of traders. Econometrica 32:39–50
4. Aumann RJ, Myerson RB (1988) Endogenous formation of links between players and coalitions: An application of the

- Shapley value. In: Roth A (ed) *The Shapley value*. Cambridge University Press, Cambridge, pp 175–191
5. Bala V, Goyal S (2000) A noncooperative model of network formation. *Econometrica* 68:1181–1229
  6. Banerjee S, Konishi H, Sonmez T (2001) Core in a simple coalition formation game. *Soc Choice Welf* 18:135–158
  7. Belleflamme P, Bloch F (2004) Market sharing agreements and collusive networks. *Int Econ Rev* 45:387–411
  8. Berge C (2001) *The theory of graphs*. Dover, Mineola (reprint of the translated French edition published by Dunod, Paris, 1958)
  9. Bhattacharya A (2005) *Stable and efficient networks with far-sighted players: The largest consistent set*. Typescript, University of York
  10. Bloch F (1995) Endogenous structures of association in oligopolies. *Rand J Econ* 26:537–556
  11. Bloch F (2005) Group and network formation in industrial organization: A survey. In: Demange G, Wooders M (eds) *Group formation in economics: Networks, clubs, and coalitions*. Cambridge University Press, Cambridge, pp 335–353
  12. Bloch F, Genicot G, Ray D (2008) Informal insurance in social networks. *J Econ Theory*. doi:10.1016/j.jet.2008.01.008
  13. Blume L (1993) The statistical mechanics of strategic interaction. *Games Econ Behav* 5:387–424
  14. Bogomolnaia A, Jackson MO (2002) The stability of hedonic coalition structures. *Games Econ Behav* 38:201–230
  15. Bollobas B (1998) *Modern graph theory*. Springer, New York
  16. Boorman SA (1975) A combinatorial optimization model for transmission of job information through contact networks. *Bell J Econ* 6:216–249
  17. Bramoulle Y, Kranton R (2007) Public goods in networks. *J Econ Theory* 135:478–494
  18. Bramoulle Y, Kranton R (2007) Risk-sharing networks. *J Econ Behav Organ* 64:275–294
  19. Calvo-Armengol A (2004) Job contact networks. *J Econ Theory* 115:191–206
  20. Calvo-Armengol A, Ballester C, Zenou Y (2006) Who's who in networks. Wanted: The key player. *Econometrica* 75:1403–1418
  21. Calvo-Armengol A, Jackson MO (2004) The effects of social networks on employment and inequality. *Am Econ Rev* 94:426–454
  22. Calvo-Armengol A, Jackson MO (2007) Social networks in labor markets: Wage and employment dynamics and inequality. *J Econ Theory* 132:27–46
  23. Casella A, Rauch J (2002) Anonymous market and group ties in international trade. *J Int Econ* 58:19–47
  24. Casella A, Rauch J (2003) Overcoming informational barriers in international resource allocations: Prices and ties. *Econ J* 113:21–42
  25. Chvatal V, Lovasz L (1972) Every directed graph has a semi-kernel. In: *Hypergraph Seminar, Lecture Notes in Mathematics*, vol 411. Springer, Berlin
  26. Chwe M (1994) Farsighted coalitional stability. *J Econ Theory* 63:299–325
  27. Chwe M (2000) Communication and coordination in social networks. *Rev Econ Stud* 67:1–16
  28. Corominas-Bosch M (2004) Bargaining in a network of buyers and sellers. *J Econ Theory* 115:35–77
  29. Currarini S (2007) Group stability of hierarchies in games with spillovers. *Math Soc Sci* 54:187–202
  30. Currarini S, Morelli M (2000) Network formation with sequential demands. *Rev Econ Des* 5:229–249
  31. Debreu G, Scarf H (1963) A limit theorem on the core of an economy. *Int Econ Rev* 4:235–246
  32. Demange G, Henreit D (1991) Sustainable oligopolies. *J Econ Theory* 54:417–428
  33. Demange G (1994) Intermediate preferences and stable coalition structures. *J Math Econ* 23:45–48
  34. Demange G (2004) On group stability and hierarchies in networks. *J Political Econ* 112:754–778
  35. Deroian F, Gannon F (2005) Quality improving alliances in differentiated oligopoly. *Int J Ind Organ* 24:629–637
  36. Diamantoudi E, Xue L (2003) Farsighted stability in hedonic games. *Soc Choice Welf* 21:39–61
  37. Durlauf S (1997) Statistical mechanics approaches to socioeconomic behavior. In: Arthur WB, Durlauf S, Lane DA (eds) *The economy as an evolving complex system II*. Addison-Wesley, Reading, pp 81–104
  38. Dutta B, Ghosal S, Ray D (2005) Farsighted network formation. *J Econ Theory* 122:143–164
  39. Dutta B, Mutuswami S (1997) Stable networks. *J Econ Theory* 76:322–344
  40. Even-Dar E, Kearns M, Suri S (2007) A network formation game for bipartite exchange economies. Computer and Information Science typescript, University of Pennsylvania
  41. Furusawa T, Konishi H (2007) Free trade networks. *J Int Econ* 72:310–335
  42. Galeana-Sanchez H, Xueliang L (1998) Semikernels and  $(k, l)$ -kernels in digraphs. *SIAM J Discret Math* 11:340–346
  43. Galeotti A, Moraga-Gonzalez JL (2007) Segmentation, advertising and prices. *Int J Ind Organ*. doi:10.1016/j.jindorg.2007.11.002
  44. Gillies DB (1959) Solutions to general non-zero-sum games. In: Tucker AW, Luce RD (eds) *Contributions to the Theory of Games*, vol 4. Princeton University Press, Princeton, pp 47–85
  45. Goyal S (2005) Learning in networks. In: Demange G, Wooders M (eds) *Group formation in economics: Networks, clubs, and coalitions*. Cambridge University Press, Cambridge, pp 122–167
  46. Goyal S (2007) *Connections: An introduction to the economics of networks*. Princeton University Press, Princeton
  47. Goyal S, Joshi S (2003) Networks of collaboration in oligopoly. *Games Econ Behav* 43:57–85
  48. Goyal S, Joshi S (2006) Bilateralism and free trade. *Int Econ Rev* 47:749–778
  49. Goyal S, Moraga-Gonzalez JL (2001) R&D networks. *Rand J Econ* 32:686–707
  50. Granovetter M (1973) The strength of weak ties. *Am J Sociol* 78:1360–1380
  51. Gravel N, Thoron S (2007) Does endogenous formation of jurisdictions lead to wealth stratification? *J Econ Theory* 132:569–583
  52. Green J (1972) On the inequitable nature of core allocations. *J Econ Theory* 4:132–143
  53. Guilhaud GT (1949) *La theorie des jeux*. Economie. Appliquee 2:18
  54. Harsanyi JC (1974) An equilibrium-point interpretation of stable sets and a proposed alternative definition. *Manag Sci* 20:1472–1495
  55. Herings PJ-J, Mauleon A, Vannetelbosch V (2006) Farsightedly stable networks. Meteor Research Memorandum RM/06/041



56. Hojman D, Szeidl A (2006) Endogenous networks, social games and evolution. *Games Econ Behav* 55:112–130
57. Hollard G (2000) On the existence of a pure strategy equilibrium in group formation games. *Econ Lett* 66:283–287
58. Inarra E, Kuipers J, Olaizola N (2005) Absorbing and generalized stable sets. *Soc Choice Welf* 24:433–437
59. Jackson MO (2003) The stability and efficiency of economic and social networks. In: Dutta B, Jackson MO (eds) *Networks and groups: Models of strategic formation*. Springer, Heidelberg, pp 99–141
60. Jackson MO (2005) A survey of models of network formation: Stability and efficiency. In: Demange G, Wooders M (eds) *Group formation in economics: Networks, clubs, and coalitions*. Cambridge University Press, Cambridge, pp 11–57
61. Jackson MO, van den Nouweland A (2005) Strongly stable networks. *Games Econ Behav* 51:420–444
62. Jackson MO, Watts A (2002) The evolution of social and economic networks. *J Econ Theory* 106:265–295
63. Jackson MO, Watts A (2008) Social games: Matching and the play of finitely repeated games. *Games Econ Behav*. doi:10.1016/j.geb.2008.02.004
64. Jackson MO, Wolinsky A (1996) A strategic model of social and economic networks. *J Econ Theory* 71:44–74
65. Kalai E (2004) Large robust games. *Econometrica* 72:1631–1665
66. Kalai E, Pazner A, Schmeidler D (1976) Collective choice correspondences as admissible outcomes of social bargaining processes. *Econometrica* 44:233–240
67. Kalai E, Schmeidler D (1977) An admissible set occurring in various bargaining situations. *J Econ Theory* 14:402–411
68. Kirman A (1983) Communication in markets: A suggested approach. *Econ Lett* 12:101–108
69. Kirman A, Herreiner D, Weisbuch G (2000) Market organization and trading relationships. *Econ J* 110:411–436
70. Konishi H, Le Breton M, Weber S (1998) Equilibrium in a finite local public goods economy. *J Econ Theory* 79:224–244
71. Konishi H, Ray D (2003) Coalition formation as a dynamic process. *J Econ Theory* 110:1–41
72. Kovalenkov A, Wooders M (2001) Epsilon cores of games with limited side payments: Nonemptiness and equal treatment. *Games Econ Behav* 36:193–218
73. Kovalenkov A, Wooders M (2003) Approximate cores of games and economies with clubs. *J Econ Theory* 110:87–120
74. Kranton R, Minehart D (2000) Networks versus vertical integration. *RAND J Econ* 31:570–601
75. Kranton R, Minehart D (2001) A theory of buyer-seller networks. *Am Econ Rev* 91:485–508
76. Li S (1992) Far-sighted strong equilibrium and oligopoly. *Econ Lett* 40:39–44
77. Li S (1993) Stability of voting games. *Soc Choice Welf* 10:51–56
78. Lucas WF (1968) A game with no solution. *Bull Am Math Soc* 74:237–239
79. Luo X (2001) General systems and  $\varphi$ -stable sets – a formal analysis of socioeconomic environments. *J Math Econ* 36:95–109
80. Mariotti M, Xue L (2002) Farsightedness in coalition formation. Typescript, University of Aarhus
81. Maschler M, Peleg B (1967) The structure of the kernel of a cooperative game. *SIAM J Appl Math* 15:569–604
82. Maschler M, Peleg B, Shapley LS (1971) The kernel and bargaining set for convex games. *Int J Game Theory* 1:73–93
83. Mauleon A, Sempere-Monerris J, Vannetelbosch V (2008) Networks of knowledge among unionized firms. *Can J Econ* (to appear)
84. Mauleon A, Vannetelbosch V (2004) Farsightedness and cautiousness in coalition formation games with positive spillovers. *Theory Decis* 56:291–324
85. Monderer D, Shapley LS (1996) Potential games. *Games Econ Behav* 14:124–143
86. Montgomery J (1991) Social networks and labor market outcomes: Toward an economic analysis. *Am Econ Rev* 81:1408–1418
87. Mutuswami S, Winter E (2002) Subscription mechanisms for network formation. *J Econ Theory* 106:242–264
88. Myerson RB (1977) Graphs and cooperation in games. *Math Oper Res* 2:225–229
89. Page FH Jr, Kamat S (2005) Farsighted stability in network formation. In: Demange G, Wooders M (eds) *Group formation in economics: Networks, clubs, and coalitions*. Cambridge University Press, Cambridge, pp 89–121
90. Page FH Jr, Wooders M (1996) The partnered core and the partnered competitive equilibrium. *Econ Lett* 52:143–152
91. Page FH Jr, Wooders M (2005) Strategic basins of attraction, the farsighted core, and network formation games. FEEM Working Paper 36.05
92. Page FH Jr, Wooders M (2007) Club networks with multiple memberships and noncooperative stability. Indiana University, Department of Economics typescript (paper presented at the Conference in Honor of Ehud Kalai, 16–18 December, 2007)
93. Page FH Jr, Wooders M (2008) Strategic basins of attraction, the path dominance core, and network formation games. *Games Econ Behav*. doi:10.1016/j.geb.2008.05.003
94. Page FH Jr, Wooders M, Kamat S (2005) Networks and farsighted stability. *J Econ Theory* 120:257–269
95. Qin C-Z (1993) A conjecture of Shapley and Shubik on competitive outcomes in the cores of NTU market games. *Int J Game Theory* 22:335–344
96. Qin C-Z (1994) The inner core of an N-person game. *Games Econ Behav* 6:431–444
97. Qin C-Z (1996) Endogenous Formations of Cooperation Structures. *J Econ Theory* 69:218–226
98. Rees A (1966) Information networks in labor markets. *Am Econ Rev* 56:218–226
99. Reny PJ, Wooders M (1996) The partnered core of a game without side payments. *J Econ Theory* 70:298–311
100. Richardson M (1953) Solutions of irreflexive relations. *Ann Math* 58:573–590
101. Rockafellar RT (1984) *Network flows and monotropic optimization*. Wiley, New York
102. Rosenthal RW (1973) A class of games possessing pure-strategy Nash equilibria. *Int J Game Theory* 2:65–67
103. Roth AE (1975) A lattice fixed-point theorem with constraints. *Bull Am Math Soc* 81:136–138
104. Roth AE (1977) A fixed-point approach to stability in cooperative games. In: Karamardian S (ed) *Fixed points: Algorithms and applications*. Academic Press, New York
105. Roughgarden T (2005) *Selfish routing and the price of anarchy*. MIT Press, Cambridge



106. Scarf H (1967) The core of an  $N$ -person game. *Econometrica* 35:50–69
107. Schwartz T (1974) Notes on the abstract theory of collective choice. Carnegie-Mellon University, School of Urban and Public Affairs typescript
108. Shapley LS, Shubik M (1969) On market games. *J Econ Theory* 1:9–25
109. Shenoy PP (1980) A dynamic solution concept for abstract games. *J Optim Theory Appl* 32:151–169
110. Shubik M (1971) The “bridge game” economy: An example of indivisibilities. *J Political Econ* 79:909–912
111. Skyrms B, Pemantle R (2000) A dynamic model of social network formation. *Proc Nat Acad Sci* 97:9340–9346
112. Slikker M, Dutta B, van den Nouweland A, Tijs S (2000) Potential maximizers and network formation. *Math Soc Sci* 39:55–70
113. Slikker M, van den Nouweland A (2001) Social and economic networks in cooperative game theory. Kluwer, Boston
114. Slikker M, van den Nouweland A (2002) Network formation, costs, and potential games. In: Borm P, Peters H (eds) *Chapters in game theory*. Kluwer, Boston, pp 223–246
115. Tardos E, Wexler T (2007) Network formation games and the potential function method. In: Nisan N, Roughgarden T, Tardos E, Vazirani V (eds) *Algorithmic game theory*. Cambridge University Press, Cambridge, pp 487–516
116. Tesfatsion L (1997) A trade network game with endogenous partner selection. In: Amman HM, Rustem B, Whinston AB (eds) *Computational approaches to economic problems*. Kluwer, Boston, pp 249–269
117. Tesfatsion L (1998) Preferential partner selection in evolutionary labor markets: A study in agent-based computational economics. In: Porto VW, Saravanan N, Waagen D, Eiben AE (eds) *Evolutionary programming VII. Proceedings of the seventh annual conference on evolutionary programming*. Springer, Berlin, pp 15–24
118. Topa G (2001) Social interactions, local spillovers, and unemployment. *Rev Econ Stud* 68:261–295
119. van Deemen AMA (1991) A note on generalized stable set. *Soc Choice Welf* 8:255–260
120. van den Nouweland A (2005) Models of network formation in cooperative games. In: Demange G, Wooders M (eds) *Group formation in economics: Networks, clubs, and coalitions*. Cambridge University Press, Cambridge, pp 58–88
121. Vega-Redondo F (2007) *Complex social networks*. Cambridge University Press, Cambridge
122. von Neumann J, Morgenstern O (1944) *Theory of games and economic behavior*. Princeton University Press, Princeton
123. Wang P, Watts A (2006) Formation of buyer-seller trade networks in a quality differentiated product market. *Can J Econ* 39:971–1004
124. Watts A (2001) A dynamic model of network formation. *Games Econ Behav* 34:331–341
125. Wooders M (1983) The epsilon core of a large replica game. *J Math Econ* 11:277–300
126. Wooders M (2008) Competitive markets and market games. *Rev Econ Design* (forthcoming)
127. Wooders M (2008) Small group effectiveness, per capita boundedness and nonemptiness of approximate cores. *J Math Econ*. doi:10.1016/j.jmateco.2007.06.006
128. Wooders M, Cartwright C, Selten R (2006) Behavioral conformity in games with many players. *Games Econ Behav* 57:347–360
129. Xue L (1998) Coalitional stability under perfect foresight. *Econ Theory* 11:603–627
130. Xue L (2000) Negotiation-proof Nash equilibrium. *Int J Game Theory* 29:339–357
131. Zissimos B (2005) Why are free trade agreements regional? FEEM Working Paper 67-07

---

## Networks: Structure and Dynamics

ERZSÉBET RAVASZ REGAN

Department of Medicine, Beth Israel Deaconess Medical Center, Boston, USA

### Article Outline

[Glossary](#)

[Definition and Relevance](#)

[Introduction](#)

[Structural Properties of Complex Networks](#)

[Dynamics on Complex Networks](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Simple graph or network** A group of  $N$  nodes (vertices) among which there exist  $L$  undirected connections (links, edges), identical in strength.

**Directed graph** A group of nodes among which connections are directed.

**Weighted network** A group of nodes among which connections are not identical in strength, but carry a weight.

**Bipartite network** A network with more than one type of node, in which connections only exist between different node types (the definition can be relaxed to a network were most, but not all links run between vertices of different types).

**Adjacency matrix  $A$**  An  $N \times N$  matrix representing the network, whose elements  $a_{ij}$  are equal to 1 when there is a link from node  $i$  to  $j$ , zero otherwise.

**Degree distribution  $P(k)$**  The probability that a node of a network, chosen uniformly at random, has degree  $k$ .

**Scale-free network** A network in which the tail of the degree distribution follows a power law (strictly speaking, the term scale-free implies  $P(k) \sim k^{-\gamma}$ , however, it is often used for networks where the tail of the distribution follows a power-law).

**Degree exponent  $\gamma$**  The power law exponent of the (tail of the) degree distribution

**Scale-free model** A growing network model proposed by Barabási and Albert [15]. The model builds a simple graph starting from a small connected group of nodes, to which new nodes are added one by one. These new nodes connect to  $m$  old nodes with probabilities that increase linearly with the degree of the old nodes.

**Shortest path (geodesic path)** The smallest collection of links that form a path through the network from one vertex to another.

**Diameter  $D$**  The length of the largest geodesic path in a network.

**Small-world network** A network in which the average shortest path length grows logarithmically (or slower) with  $N$ .

**Node betweenness (betweenness centrality or load)**

The number of shortest paths between nodes of the network that run through a given node [62].

**Edge betweenness** The number of shortest paths between nodes of the network that run through a given edge.

**Clustering coefficient  $C$**  The fraction of connections that are realized between the neighbors of a node:

$$C_i = \frac{2 n_i}{k_i (k_i - 1)},$$

where  $n_i$  denotes the number of links connecting the  $k_i$  neighbors of node  $i$ . (The average clustering coefficient is given by  $\langle C \rangle = \frac{1}{N} \sum_i C_i$ . An alternative global measure of clustering, also called *transitivity*, is the fraction of node triples that are linked into triangles.)

**Assortativity coefficient** A measure of the tendency of links to run among nodes that are similar in some respect. If the similarity is described by a scalar quantity (most often the node's degree), then the assortativity coefficient is given by

$$r = \frac{\sum_{x,y} xy (e_{x,y} - a_x b_y)}{\sigma_a \sigma_b},$$

where  $x$  ( $y$ ) is the scalar at the origin (end) of a link,  $e_{x,y}$  denotes the fraction of all edges in the network that go from nodes with value  $x$  to ones with value  $y$ ,  $a_x$  ( $b_y$ ) is the fraction of edges that start (end) at a link with values  $x$  ( $y$ ), and  $\sigma_a$  ( $\sigma_b$ ) is the standard deviations of the distributions of  $a_x$  ( $b_y$ ) values [107].

**Modularity  $Q$**  The number of links between nodes within the same community minus the number expected by

chance:

$$Q = \frac{1}{2L} \sum_{i=1}^N \sum_{j=1}^N (A_{ij} - P_{ij}) \delta_{g_i, g_j},$$

where node  $i$  ( $j$ ) belongs to the community  $g_i$  ( $g_j$ ).  $P_{ij}$  gives the expected number of links between two nodes if the network is random with respect to communities [110]. In the simplest case, in which the null model is a random network,  $P_{ij} = 2L/N^2$ . A more suitable assumption is  $P_{ij} = k_i k_j / 2L$ , which preserves the degree distribution of the network in question (the expected degree of node  $i$  is  $\sum_j P_{ij} = k_i$ ) [109].

## Definition and Relevance

Scientific research has had a long history of bottom-up approaches, which break the system into small or elementary constituents and map out interactions between these components. The Standard Model describing elementary particles and the four types of interactions governing our world is perhaps the most successful example. Biology has developed a very detailed description of cellular components such as the DNA molecule or the various proteins and metabolites. Furthermore, many of the interactions that govern a cell's life have been investigated in great detail, but mainly in isolation: transcription of DNA, protein assembly, enzyme function, etc. Perhaps not surprisingly, the first attempts to understand complexity in physics were focused on small, simple system with complex dynamics: chaos theory. Nonetheless, large natural or social systems, like a cell, an ecosystem or the Internet are much more intuitive examples of complex systems. A meaningful description of these systems requires more than a mere account of the constituent parts: one does not understand the way the Internet works by detailing the physical characteristics of computers. Nor is the sequence of a cell's genome the final tool for understanding its behavior.

Complex systems display characteristics that are fundamentally determined by their organization, emergent phenomena created by all the interacting constituents. In many cases, if one takes a step back, avoiding the details of the interactions, a complex system as a whole is made up of an assemblage of generic elements and connections; in other words, it looks like a network. For example, a cell's metabolism is maintained by a biochemical network, whose nodes are substrates and links chemical reactions [75]. Equally complex webs describe human societies, whose nodes are individuals and links represent social interactions [147], the World Wide Web (WWW) [6,28], where nodes are Web documents con-

nected by URL links, the scientific literature, whose nodes are publications and links citations [44,124,125], or language, made of words linked by various syntactic or grammatical relationships [52,61,129]. Due to the diversity and large number of the nodes and interactions, the system-level characteristics of these networks remained largely unknown and unexplored prior to the last decade. At the same time, the inability of contemporary science to address the properties of complex networks limited advances in many disciplines, including molecular biology, computer science, ecology and the social sciences. The recent availability of system-level data on the network of interactions in large numbers of systems has opened the door for interdisciplinary research in fields where the behavior of the system as a whole is a central question. Recognizing generic organizational principles and order behind diversity and apparent randomness in these different systems has certainly been a surprise along the way.

## Introduction

The explosion of available data describing interaction of hundreds to millions of components in systems like the Internet or the protein interaction network is a recent development for the study of complex systems. Nonetheless, networks are not new to mathematics or the social sciences. Graph theory was born from the famous Königsberg bridge problem and its even more famous solution by Euler in 1736 [59]. The problem was simple: find a closed walk that visits each of Königsberg's seven bridges once, but only once. The trouble was, nobody could find such a walk. Euler drew a four-node graph of the pieces of land connected by the bridges and showed that graphs that have nodes with odd degrees cannot have closed self-avoiding paths. (All four pieces of land in Königsberg are connected via an odd number of bridges). Euler's emphasis on the topology of the bridge problem as key to the solution marks the starting point of two prolific subfields of mathematics: topology and graph theory.

The early 1920's witnessed the birth of social network analysis, a subfield of sociology trying to understand how social interactions organize. Data gathering methods limited the size of the networks studied, nonetheless many network measures important today were defined: degree distribution, betweenness, clustering and the small world effect.

In the 1950's two prolific hungarian mathematicians, Erdős and Rényi, took the challenge of describing the structure of large social networks from social science to mathematics and formulated their famous random graph model [57,58]. Their key innovation was a statistical ap-

proach to graph theory: their theorems were proved on the set of all networks generated by the rules of their model. Their random network rules were simple: take  $N$  nodes and connect each pair with probability  $p$ . They found that above a certain threshold probability the ensemble of random graphs undergoes a percolation-type phase transition from a graph made of small disjoint subgraphs to one with a giant component comparable with the system size. Random graphs have a Poissonian degree distribution: they are homogeneous networks with a well-defined average degree. Erdős and Rényi pointed out that random networks are "small worlds": their diameter scales with the logarithm of system size (node number), quite different from the power-law relationship that holds for regular lattices.

The Erdős-Rényi model guided our thinking about complex networks until the end of the 1990's, when two seminal papers triggered a very rapid growth of the field which today is called complex networks research. The first paper, published by Watts and Strogatz in 1998 [148], showed that natural systems such as the neural network of the *C. Elegans* worm, the power grid and the network of movie actors connected by feature films have a topology somewhere between regular lattices and random graphs. These networks have large clustering coefficients, but also the small world property. The small-world model presented in the paper is the first important step away from the random world of the Erdős-Rényi graph. The second paper, published one year later by Barabási and Albert, showed that the degree distributions of the movie actor network, the WWW and the power grid are not Poissonian: they have a power-law tail, inspiring the term scale-free networks [15]. The list of power-law tailed degree distributions measured on natural and man-made systems is still growing, with degree exponents that rarely fall outside the (2,3) interval. Power law degree distributions tell us that most real networks are highly heterogeneous: the majority of nodes have very small degree, but a few hubs with degrees orders of magnitude larger than the average also exist, along with nodes with degrees of all scales in between. A good example is the US airport network connected by direct flights: Chicago and Atlanta at the high end of the degree scale, the numerous regional airports at the low end.

Naturally, the sudden explosion of data and tools to explore them brought us closer to the heart of questions fundamental to understanding complex systems: what drives their organization, what defines their emergent properties. What do these networks do, what is their function? This question is natural to biology. A metabolic network has a well-defined job in the living cell: it fuels the cell with

energy, nutrients and building blocks, and it does so adaptively and robustly. Man-made complex networks, such as the Internet or WWW, the power grid or the network of synonyms in a language also perform well-defined functions. In contrast with the delicate and well-thought-out internal structure of a computer, these systems were not designed from scratch with their function in mind: they evolved and emerged naturally. Also in contrast with the fragility of a computer (pull out a random element and it stops working), complex systems in biology and technology have a remarkable resilience to random node failure: the internet does not die when a few routers go down, in fact this probably is its normal mode of operation.

Questions about the characteristics mandated by function, common to all these systems, along with questions about the selection principles that shape these structures, are at the heart of our quest to understand complex function. These questions are typically formulated in terms of the dynamics natural to the network in question: metabolic flux driven by enzymes, packet traffic on the internet, web browsing, electric current flow on the power grid, the spread of HIV on the sexual contact network, etc. Understanding how structure affects dynamics and vice versa is in the spotlight of complex networks research today.

### Structural Properties of Complex Networks

Let us look at the metabolism of a cell as an example that highlights the increasingly detailed ways one can pose the question: what is the large-scale structure of cellular metabolism? At first glance, the metabolic network is a *simple graph* in which metabolites (the nodes) are connected by chemical reactions. This representation tells us whether the network is homogeneous in degree or has hubs, whether it is a “small world”, whether it has a community structure. A more detailed description of the system takes into account the direction of chemical reactions, since a large number of them are not reversible in a living cell. This leads to a *directed network* in which the in- and -out-degree distributions can be different and paths are directed. The next step is a *weighted network*: metabolites are characterized by their concentrations in the cell, edges are weighted by fluxes carried by reactions. A different way of adding complexity to the representation is by constructing a *bipartite graph*, where metabolites on one side connect to reactions on the other.

The above example suggests a natural organization for this section: presentation of simple graphs and their characteristics followed by more detailed networks, in parallel with specific real-world examples and models.

### Simple Graphs

**Degree Distribution** The Königsberg bridge network had only four nodes; the network examples we cite nowadays have from hundreds to millions. Most of them show a degree heterogeneity best captured by the degree distribution, which carries more information than the average degree, as pointed out by Barabási and Albert [15]. Examples of scale-free networks include networks of metabolic reactions [75], genetic regulatory interactions [95,128,138], earthquake event correlations [13], word co-occurrence and synonyms [52,61,101,129], power lines [8,148], air routes [11,18], the Internet [29,60,67], the World Wide Web [1,2,6,83,88], software systems [102], Wikipedia links [31], phone calls [4], e-mails [55], co-authorship [17,104], scientific citations [124], World trade [127], innovation flow [48], sexual partnership [90] and the list goes on.

Barabási and Albert had a much shorter list of examples to work with in 1999. Nonetheless, they saw that three very different networks, the actor network, power grid and World Wide Web had similar degree distributions, different from that of a random network model. They argued that the Erdős–Rényi random network model lacks two important features present in real-world systems. First, it is a static model, while most real-world networks constantly evolve and grow. Second, the random network model is too democratic: the probability of linking any two nodes is constant. Barabási and Albert proposed that in real networks the nodes that already have a large degree are more likely to receive new links as the network grows (think of Google in the WWW network). Preferential attachment accompanied by network growth was the first mechanism ever reported to reproduce the scale-free (power-law) degree distribution seen in real networks. Interestingly, though, Barabási and Albert were not the first to report it. A sociologist named Price reported the first example of a scale-free network back in 1965, that of citations linking scientific papers [44]. Building on the “the rich get richer” idea proposed to explain wealth distributions, Price constructed a network model for citations in which the more citations a paper has, the more likely it is to acquire further citations [45].

The scale-free model, simple and analytically solvable, propelled scale-free networks and preferential attachment to the forefront of the expanding field of complex networks research. Barabási and Albert showed that both growth and preferential attachment are essential for obtaining a scale-free network. Growth along with uniform attachment leads to an exponential degree distribution, while no growth with preferential attachment leads to

a Gaussian distribution (although the system does start out with a transient power law) [16]. They also considered the effects of random rewiring and internal link formation, showing that internal link dynamics cause deviation from the power-law at low degrees, observed in real-world networks [5]. A variety of models and analytical methods were developed in the wake of these papers, addressing the effects of further changes in rules of growth or attachment on the degree distribution. The exact solution for the degree distribution of the Barabási–Albert model was worked out by graph theorists Bollobás and Riordan [27].

One of the most important generalizations of the original scale-free model was the study of nonlinear preferential attachment by Krapivsky and Redner [85,87]. They found that power-law scaling is destroyed by nonlinearity: sublinear attachment leads to a power law multiplied by a stretched exponential, while faster than linear attachment causes the network to “condense”: the fraction of nodes connected to a single super-hub is finite in the thermodynamic limit. Indeed, the simple linear attachment rule of the scale-free model was later verified in real systems: citation networks, the Internet, the actor and scientific collaboration networks [76,103].

Variations on the scale-free model include linear preferential attachment offset by a constant [53], internal edge creation and removal [50,86], growing average degree [51] (seen in the WWW and co-authorship networks [17]), and edge rewiring [5,135]. One of the challenges of modeling the WWW with the original scale-free model was spotted by Adamic and Huberman: while the oldest nodes are the ones with the highest degree in the model, the WWW does not show this correlation [2]. Bianconi and Barabási proposed a multiplicative fitness model [21] in which the attachment rule is influenced by the degree as well as the “worth” of a node. This model generates both scale-free networks and “winner takes all” scenarios; the transition between the two outcomes maps beautifully onto a Bose-Einstein condensation [22].

The presence of preferential attachment can be justified in many real systems such as the WWW, citation, collaboration or airport networks through the larger visibility of high-degree nodes and the advantage nodes gain by linking to them. There are, however, scale-free networks where a different mechanism leading to preferential attachment is necessary. Biological networks offer intriguing examples: the metabolic and the protein-protein interaction networks are scale-free, even though their connections are governed by biochemistry and not choice. Protein interaction networks have inspired a class of models based on gene duplication (duplication of a node and all

its links) and subsequent mutation (addition and/or deletion of some of the copy’s links) [130,145]. Preferential attachment in these models is a consequence of the evolutionary dynamics: nodes with a higher degree have more duplicating neighbors, thus receiving more links. Vertex (node) copying has been proposed as a possible mechanism for the growth of the WWW [83] and auto-catalytic networks [74].

In the spirit of the Erdős–Rényi model, Bender and Canfield proposed an ensemble model for scale-free networks. The configuration model describes the group of all networks with a prescribed degree sequence [20]. If the sequence is chosen from a power-law distribution, the resulting networks are naturally scale-free. However, the average properties of the ensemble carry no other characteristics intrinsic to evolving models (such as degree correlations or clustering). Moreover, the simple definition of the model makes it ideal for analytical approaches. Newman et al. used probability generating functions to calculate exact expressions for average path length and clustering in configuration networks. (They used a generalized definition: the ensemble of all networks with an ensemble of degree sequences drawn from the same distribution.) [112]. Many analytical results for scale-free networks were proven using the configuration model: they have asymptotically vanishing clustering coefficients [42] and they are ultra-small for the degree exponents measured on most real-world networks ( $\gamma \in (2, 3)$ ): their average path length scales as  $\mathcal{O}(\log \log N)$  [34].

Some real networks are embedded in physical space and have physical connections: brain networks, the power grid, the Internet, airport networks, streets, public transportation systems, highways and rivers. Some of the above examples (brain networks, streets, rivers) are missing from the list of scale-free networks: spatial constraints can be forbidding to the formation of hubs. Several studies with both preferential attachment and bias towards shorter links show that large length costs can destroy the scale-free nature of spatial networks [19,92,151]. Systems with fixed maximum link length, such as wireless networks in which the range of a particular device is much smaller than the physical size of the system are not scale-free. Conformation networks made of all physically allowed conformations of a system (a polymer or bead chain) are also homogeneous structures naturally embedded in an  $n$ -dimensional configuration space defined by the system’s degrees of freedom [123,126].

The distribution of nodes in space can also have great influence on network topology, as shown in a detailed study of the Internet by Yook et al. [152]. They found that the router density distribution is a fractal with the



same dimension  $D = 1.5$  as the population density distribution. They were able to mimic the topology of the Internet using a simple evolving model in which nodes are distributed according to a fractal distribution and incoming nodes connect to old ones using preferential attachment divided by some power of their distance. Interestingly, the only model parameters able to reproduce the internet's topology were the ones actually measured for the real system: fractal dimension  $D = 1.5$ , linear preferential attachment and a probability of connecting two nodes that is inversely proportional to the distance between them. A similar model was also successful in describing the topology of the world-wide airport network [69].

### Paths on Networks, Small Worlds and Betweenness

A networks' most basic function always requires some type of communication along its edges. Thus it is natural that average shortest paths, network diameter and local betweenness measures are of great interest to networks research. Erdős and Rényi proved that the diameter of their random network model scales as the logarithm of node number. Watts and Strogatz called such networks small worlds [148], in tribute to the "small world phenomenon" in sociology: the idea that one can connect any two people on Earth by about six handshakes between mutual acquaintances. In 1967, an ingenious experiment by Milgram proved the existence of short paths of an average six hops between random people in the US, known as "six degrees of separation" [94]. The small world nature and searchability of small-world networks allow movie lovers to find short chains of movies connecting their favorite actor to Kevin Bacon. Mathematicians (and networks researchers) play the same game on the co-authorship network: one's shortest path to Pál Erdős is called the Erdős number.

The small world model introduced by Watts and Strogatz is based on the idea that real networks are in between random graphs and regular lattices: they are highly clustered (as a regular lattice), but they also have shortcuts. The model is built starting with a regular low-dimensional lattice to which one adds (or rewires) a certain number of edges, shortcuts between distant parts of the lattice [148]. Watts and Strogatz showed that increasing the number of shortcuts turns regular lattices into random networks, moreover, a small number of shortcuts is sufficient for the small-world effect without destroying the high clustering coefficient of the original lattice [111,148]. Most real networks display both characteristics of the small world model: short average path lengths and high clustering. However, the small world model has a peaked degree distribution, thus the topology of real networks with scale-

free degree distribution is fundamentally different from the ones generated by the model. The Barabási-Albert scale-free model, on the other hand, fails to account for the high clustering of most real networks.

Logarithmic or slower increase of the average path length was proved for a variety of network models [25,26,32]. Bollobás and Riordan showed that the average shortest path length in scale-free networks grows no faster than  $\log N / \log(\log N)$  [25]. Cohen and Havlin used the configuration model to show that random scale-free networks with  $\gamma \in (2, 3)$  are ultra-small: their average shortest path length grows as  $\log(\log N)$  [34].

The structure of shortest paths is crucial to any communication or flow between nodes of a network. Naturally, betweenness was found to be the relevant quantity when one deals with congestion or disruption of flow in networks. Goh et al. measured the distribution of node betweenness values for a variety of real networks as well as models, and showed that it follows a power-law with only two distinct exponents [66].

**Clustering and Network Motifs** Clustering is a rediscovery of "network density", a quantity widely used in sociological network analysis: it provides a measure of how well one's acquaintances know each other. High values have been observed in social networks, but also in most other real-world networks, motivating the development of clustered scale-free models. The Holme-Kim model, aimed at creating scale-free networks that also have high clustering coefficients, is a straightforward generalization of the scale-free model, with triangle-forming steps complementing preferential attachment [72]. Klemm and Eguíluz introduced a citation network model based on the idea that papers are only cited for a limited stretch of time before they are forgotten [84]. The model is built by the constant addition of nodes that connect to all "active" nodes and join their ranks. Then, one of the active nodes is deactivated with a probability inversely proportional to its in-degree (offset by a constant). This model leads to scale-free networks with high clustering coefficient, but fails to capture their small-world nature: visualized, the networks look like tubes. Gene duplication models, while motivated by evolutionary arguments, also lead to clustered scale-free networks [130,145].

An interesting generalization of the clustering coefficient was introduced by Uri Alon's group: network motifs are significantly over- or underrepresented patterns of connections between  $n$  vertices (compared to the randomly rewired network). Distinct characteristic motifs were found in regulatory networks, food webs, neural networks and WWW, corresponding to local functions

performed by the network [95,128]. For example, different types of feed-forward loops (FFLs, directed motifs of 3 nodes:  $A \rightarrow B$ ,  $B \rightarrow C$  and  $A \rightarrow C$ ) in genetic regulatory networks perform distinct signal processing roles as shown by both simulation and experiments with living cells [10]. FFLs made of activating interactions only (abundant in the regulatory network) can filter out transient changes in the concentration of the input node, while also delaying the turn-on or turn-off of the output node. On the other hand, FFLs where the  $B \rightarrow C$  link is a repressing one (also abundant) act as pulse generators.

**Degree Correlations and Mixing Patterns** In social networks, links between people who are alike are more common; popular people are connected with popular people. Assortative mixing means that degrees at two ends of an edge are correlated, as measured by the conditional probability,  $P(k'|k)$ , that a node with  $k$  links is connected to another one with  $k'$  links. Measurements on the Internet and protein interaction networks show that in these systems, as opposed to their social counterparts, small degrees are more likely to connect to high degree nodes [93]. The conditional probability is difficult to measure in most real networks due to poor statistics, although it is convenient in analytical work. A more compact representation of degree correlations was defined by Pastor-Satorras et al. [119]: the average degree of neighbors as a function of degree,  $k_{nn}(k)$ , which decreases with  $k$  for disassortative systems. One can further simplify the measure of assortativity by calculating the Pearson correlation coefficient of degrees at the ends of a network's edges: the assortativity coefficient [105,107]. Intriguingly, most social systems are assortative (actor network, company directors, coauthorship networks, phone calls, email address books), while most technological and biological systems prefer disassortative mixing (WWW, Internet, train routes, software packages, software classes, electronic circuits, peer-to-peer networks, metabolic networks, food webs, neural networks).

Degree is not the only property of a network that can show assortative or disassortative mixing. In networks where nodes can be classified in types of some kind, mixing between types can also be characterized by the assortativity coefficient (see definition). Maslov et al. showed that there are three main types of nodes in the Internet: high level providers who manage the backbone and trunk lines, Internet Service Providers who bring the network out to end-users and the end-users themselves. These three types show strong disassortativity: few end-user to end-user or ISP to ISP links in the network. In social networks, mixing by race, age or income has been observed [107].

**Communities, Hierarchy and Fractality** Community structure is an intuitive feature of complex networks: one expects them in social systems (circles of friends), biological networks (functional units), the WWW (websites related to a topic or organizations), co-authorship networks (scientific fields and sub-fields), etc. Communities within networks are structural features related to the function of the network as a whole, and are thus expected to have a strong influence on their dynamics. Defining and finding network communities has its history in both sociology and computer science, and has been revisited many times. The methods developed along the way paint a rich picture of the structural diversity of complex networks.

*Non-Overlapping Community Structure* Hierarchical clustering or cluster analysis is widely used in the study of social networks. The first step in hierarchical clustering is the construction of a similarity measure between network nodes. Next, each node is assigned to a separate cluster (the leaves of the dendrogram). The two most similar clusters are joined to form junctions of the dendrogram, until all clusters have been united, forming the root. Many similarity measures have been defined and used successfully. Structural equivalence, used in social network analysis, gives two nodes the highest similarity score if they have the same pattern of relationships. It can be measured using Euclidean distance or Pearson correlation coefficient between rows of the adjacency matrix, as well as topological overlap, defined as the number of overlapping neighbors divided by the smaller of the two degrees [122].

Girvan and Newman introduced a famous community detection algorithm that is similar to hierarchical clustering, but works divisively [64]. It is based on the idea that the edges most likely to run between communities are the ones that have the highest edge betweenness centrality. They first remove the edge with the highest betweenness, then recompute edge betweenness and keep removing edges until the network falls apart into non-connected nodes. As the network falls apart, they draw a dendrogram where each joint is a splitting event.

Hierarchical clustering works with varying success for different systems, but it cannot determine how many communities there are in the network. Newman et al. introduced an elegant measure for the "quality" of any given partitioning into communities, called modularity, which compares a community partitioning to a null model that can be appropriately chosen for the system at hand [109]. The Girvan–Newman algorithm, together with the dendrogram cut that maximizes modularity, has been successfully used for many different social and biological networks [12,24,71,110,141,150]. The main drawback of the

method is computation time ( $\mathcal{O}(N^3)$  on a sparse graph): it is not feasible for networks with more than a few thousand nodes.

In a quest to develop a faster community detection method closely tied to the definition of communities, Newman used direct optimization of the modularity measure [108]. Since finding the best partitioning is an NP-hard problem, his first approach was a greedy optimization. All nodes start out as separate communities, which are then repeatedly joined such that the increase in modularity is maximal (or decrease is minimal). Newman showed that the method works well both in tests and real networks, and it works quite fast:  $\mathcal{O}(N^2)$  time for sparse graphs. Still not fast enough? A collaboration with Clauset and Moore resulted in an algorithm that performed the same optimization in  $\mathcal{O}(N \log^2 N)$  for sparse graphs with many levels of communities [33], solving the problem of partitioning networks with billions of nodes. (The general result for runtime is  $\mathcal{O}(KD \log N)$ , where  $K$  is the total number of links and  $D$  is the depth of the generated dendrogram).

Interested in the theoretical foundations of community structure detection and its relationship to matrix spectra, Newman argued that community detection requires the use of the modularity matrix [109] in place of the Laplacian matrix, as done by traditional spectral methods of graph partitioning. (The Laplacian matrix of a graph,  $L$ , is a real symmetric matrix with elements  $L_{ij} = k_i \delta_{ij} - A_{ij}$ .) The modularity matrix is defined as  $B_{ij} = A_{ij} - P_{ij}$  (see modularity definition). He showed that the eigenvalues and eigenvectors of the modularity matrix encode the networks' community structure. Methods based on this relationship perform as well, if not better, than previous ones, but more importantly, they are also able to detect "anti-modular", or bipartite structure [109].

*Hierarchical Community Structure and Fractal Networks*  
Ravasz et al. argued that in many real networks with interesting modular structure there is no ideal partitioning into distinct modules: the network is built of small, very cohesive communities, hierarchically embedded in larger, less cohesive ones [121,122]. Their hierarchical model is a deterministic construction of a scale-free network with hierarchically embedded modules. They showed that hubs on all scales unite communities on all scales: small nodes are part of small, very cohesive clusters (thus have large clustering coefficients), larger nodes serve as connectors of these clusters, while the largest hubs span modules at the highest level of organization (and have low clustering coefficients). Indeed, they found that the clustering co-

efficient  $C(k)$  decreases with increasing  $k$  in a variety of real networks, such as the metabolic network, synonyms, movie actors, the WWW and the Internet represented at Autonomous System level (each node is a domain, not just a computer), and it often scales as  $1/k$  [121]. This scaling was found in a variety other networks: software systems [102], the World Trade network [127], the worldwide airport network, the co-authorship network of <http://arxiv.org/archive/cond-mat> [18,104] and protein folding networks [120]. Not all networks are hierarchical: the power grid and the Internet at router level show no scaling of the clustering coefficient [121]. Ravasz et al. suggested that spatial embedding of both systems, where length costs are significant, works against hierarchical modularity. Both networks distribute something to a physical area. Thus, the formation of tight communities is not required for these networks to function.

Hierarchical organization in metabolic networks was found to reflect biological function on different scales of organization: the branches of the dendrogram obtained by hierarchical clustering (using topological overlap as similarity measure) correspond to known functional classes of the metabolism on two different levels of organization [122].

The question of self-similarity or fractal nature has been on the mind of network researchers since the scale-free degree distribution was observed. Are networks like fractals, self-similar on all scales? Hierarchical organization strengthens this image: most networks not only have degrees in a broad range of scales, they are also made of modules of different scales embedded into each other. The problem with complex networks as fractals was, of course, that most networks of interest to us are also small-world [148]. In a small-world network the number nodes at a distance  $l$  from any given node increases exponentially. One expects this number to grow as a power law in the case of a fractal object: measuring the "mass" within distance  $l$  from a point on the object, called the cluster growing method, is one way of measuring fractal dimension.

While some deterministic models have been constructed with the idea of fractality in mind [54,77], the breakthrough in understanding topological self-similarity in complex networks was brought forth by Song, Havlin and Makse [132,133]. They generalized the standard box counting method for measuring fractal dimension of a physical object to complex networks. How does one cover a network with boxes of different size? Divide all nodes in groups such that the shortest path between any two nodes in a group is at most  $l_B$  long: these are the boxes. Use the smallest number of groups necessary to do this

(or a decent approximation) and repeat it for  $l_B \in [2, D]$ . The results of this procedure were quite surprising: many scale-free real-world networks, such as the WWW, actor network and various metabolic networks show a neat fractal scaling between the number of boxes and their size. Thus these networks are self-similar, fractal structures. To prove their point further, the authors used a renormalization procedure where they collapsed each box into a node and linked these new nodes to each other if any member of the original boxes had a connection. The networks renormalized this way were also a scale-free, with the same degree distribution exponent, independent of the box size used for renormalization [132].

The question of how these networks are small world and self similar at the same time has also been resolved: the two methods of determining the fractal dimension of an object, box counting and cluster growing are not equivalent on complex networks with broad degree distribution. While box counting covers all the hubs only once (they are assigned to one box only), cluster growing finds the hubs for almost any choice of seed node, thus bringing a large part of the network into the cluster with them. This explains the exponential increase in the number of nodes within a distance  $l$ , and shows the small-world property of the system. Box counting, on the other hand, can reveal the fractal nature of the network, if present [132].

Song et al. uncover some of the requirements of fractality via proposing a network growth model based on reverse renormalization. Starting from one point, the network grows by nodes transforming into small clusters in each iteration. The conversion from a node to a small cluster mimics the renormalization process in reverse: the degree of a node grows by multiplication with a scaling factor (thus nodes of previous iterations become the hubs). The newly formed clusters have a diameter of  $b_B$  (the box size). They showed that the key feature that influences fractality in the emerging networks is how the clusters connect: if the link always runs between the hubs (leading to assortative mixing), the result is small-world network that is not a fractal. The internet at router level, found to be non-hierarchical, is a good example of such an assortative, non-fractal network. On the other hand, if the clusters only connect to each other via the non-hub (newly created) nodes, the resulting network is a fractal, but it loses its small-world character. Many real-world networks, however, seem to be both fractal and small-world (WWW, actor network, protein interaction networks, metabolic networks). Indeed a very small number of hub-to-hub connections in this model can restore the small-world property of the network while preserving its fractal nature.

**Overlapping Community Structure** The community detection methods presented thus far assign each node to only one community. Palla et al. argued that many nodes in real networks belong to more than one community [114]: proteins can simultaneously be part of several complexes, people have disjoint groups of acquaintances from friends to work to extended family. They proposed a community definition that allows them to capture overlap between communities. Their method is based on “ $k$ -clique rolling”: a  $k$ -clique community is the union of all  $k$ -cliques (complete subgraphs of size  $k$ ) that can be connected through a series of  $k$ -cliques that share  $k - 1$  nodes. Increasing values of  $k$  lead to smaller, denser, but more disjoint communities. They find that  $k$ -clique communities in real networks (co-authorship, word association and protein interaction networks) have a power-law size distribution, and there are significant overlaps: the size distributions of overlap as well as node membership have fat tails.

### Directed Networks

Directionality of a link is often important in real networks: scientific citations, url links, genetic regulatory interactions are only a few examples of inherently asymmetric connections. Directed links require the separate measurement of in- and out-degree distributions. In case of the WWW both are power-laws, but the degree exponents differ:  $P_{in}^{WWW}(k) \sim k^{-2.1}$  and  $P_{out}^{WWW}(k) \sim k^{-2.45}$  [6]. Genetic regulatory networks show an even stronger difference. Their out-degree distributions are power laws, however, the in-degree ones are scaled: genes cannot receive input from hundreds of transcription factors.

Another interesting consequence of directed links is the rich structure of directed paths. They were found to partition the WWW as well as metabolic networks into parts resembling a bow-tie [28,91]: a strongly connected component, in which there is a directed path in both directions between any pair of nodes, an IN-component, the nodes of which can reach the strongly connected component but cannot themselves be reached, and an OUT-component that can be reached from, but has no directed paths leading into the strongly connected component. Not all directed networks have bow-tie structure: some of them are entirely acyclic, with no directed loops. Citation networks are a natural example: one can only cite papers already published. However, genetic regulatory networks are also acyclic (not considering auto-regulatory loops), even though the cause of this is now yet understood [14,138].

Gradient networks are directed graphs that generalize the concept of gradients from continuous scalar fields

to networks. They capture the backbone of a gradient-induced flow on complex networks: given a substrate graph with a scalar value associated to every node, its gradient network is formed by the collection of all directed links that lead from every node to its neighbor with the highest scalar value [139]. These directed links form collections of trees. For an independent, identically distributed association of random variables to the nodes of an Erdős–Rényi graph, the generated gradient network has been shown to be scale-free, with a connectivity exponent of  $\gamma = -1$ . This finding can be proved for any substrate graph with no loops shorter than 5 [140]. Moreover, the gradient networks' scale-free nature seems to be universal: it was numerically observed for a wide variety of substrate networks (including ones with short loops): regular and random trees, Erdős–Rényi and small-world networks, high dimensional regular lattices and  $n$ -tori, random geometric networks [39], the scale-free and the configuration model [123,140].

### Weighted Networks

Real world networks display significant heterogeneity in the strength of their connections. The distribution of link weights has been found to have a heavy tail in metabolic networks: the steady-state flux distribution follows  $\Theta(w) \sim (w_0 + w)^{-1.5}$  ( $w_0 = 3 \cdot 10^{-4}$ ) [9]. Moreover, the average link strength scales with the connectivity of the nodes at the two ends as  $\langle w_{ij} \rangle \sim (k_i k_j)^{0.5}$  in both metabolic networks and the Worldwide Airport Network, indicating correlations between node degree and links weights [9,18]. Scientific collaboration networks, on the other hand, have a weight distribution that is not correlated with degree.

The existence of weighted links requires a generalization of most network measures:

- **Node strength and strength distribution.** Node strength,  $s_i$  is a natural generalization of the node degree,  $k_i$ :

$$s_i = \sum_{j=1}^N w_{ij} A_{ij} ,$$

the sum of weights over the links of node  $i$ . The strength distribution,  $P(s)$ , is typically also heavy tailed [18].

Weights are often dependent upon topology (see metabolic and airport networks), expressed by a non-linear relationship between node strength and degree:  $s \sim k^\beta$  (in the airport network  $\beta = 1.5$ ). Heterogene-

ity in the weights around a node can be measured using

$$Y_i = \sum_{j=1}^N \left( \frac{w_{ij}}{s_i} \right)^2 A_{ij} .$$

If all edges have the same weight,  $Y(k)$  scales as  $1/k$ , while if one weight is significantly larger than the others,  $Y(k) \simeq 1$ . In metabolic networks  $Y(k)$  was found to scale as  $k^{-0.27}$ , indicating that metabolites used in a large number of reactions are more likely to have one high-flux reaction dominating their production (consumption) [9].

- **Weighted clustering coefficient.** Defined by Barrat et al. as

$$C_i^W = \frac{1}{s_i (k_i - 1)} \sum_{j=1}^N \sum_{m=1}^N \frac{w_{ij} + w_{im}}{2} A_{jm} A_{ij} A_{im} ,$$

the weighted clustering coefficient is often compared to the standard one:  $C^W > C$  means that the triangles of the network are preferentially formed by high-weight links. This is indeed the case in the co-authorship network, for nodes with  $k > 10$ : more established investigators form stable, high-weight cliques (research groups) from which the main volume of their publications originate. A similar phenomenon can be seen in the world-wide airport network: larger airports form high passenger-flux triangles, so called “rich-clubs” [18].

- **Degree Correlations.** The weighted average degree of neighbors is defined as:

$$k_{nn,i}^W = \frac{1}{s_i} \sum_{j=1}^N w_{ij} k_j A_{ij} ,$$

a sensitive probe into the structure of weighted networks. The behavior of  $k_{nn}(k)$  in the airport network shows a plateau for airports with more than 10 direct flights, indicating no degree preference. However,  $k_{nn}^W(k)$  increases in a wider range and is in general larger  $k_{nn}(k)$ , showing that while large airports do not preferentially connect to large airports, the high traffic links run among members of the “rich-clubs” [18].

### Dynamics on Complex Networks

The main theme of this section is: how does the structure of a network influence its dynamics? Most lessons of statistical mechanics are worth revisiting when the underlying space is a network: does an inhomogeneous topology change the dynamics? The answer, as we will see, is yes, in interesting ways.



## Robustness and Vulnerability of Complex Networks

Perhaps the most intriguing feature of complex systems is their robustness. Close to 75% of the genes in *E. coli* are non-essential: the organism can survive without them (under one set of growth conditions) [63]. If the webpage of a company or university goes down, the WWW as a whole is still perfectly functional, the effect of the failure is local. On the other hand, an intentional “attack”, similar to the denial-of-service attacks that crippled Yahoo, Amazon, eBay, CNN and a few other very popular websites in February of 2000, can substantially cripple the function of a complex system.

Vulnerability and robustness go hand in hand in most complex systems: bad weather in Atlanta (although usually not labeled an “attack”) has very different consequences for air traffic than problems at a regional airport. Opinions expressed in the New York Times are much more likely to spread and influence events than opinions in a small town local newspaper. These trivial examples hint at interesting consequences of in-homogeneous network topology, brought to light in a 2000 Nature paper by Albert, Jeong and Barabási [7]: while scale-free networks are largely unaffected by random failure, they are very sensitive to change in their highly connected, central nodes, rendering them vulnerable (and/or responsive) to planned, targeted interventions.

**Resilience** One of the simplest indicators of robustness under the damage done by node or link removal is a structural one: the size of the largest connected component. Numerical studies performed by Albert et al. show that the giant connected component of Erdős–Rényi random networks falls apart at a much lower fraction of randomly removed nodes than of a scale-free network [7]. A fraction of nodes the removal of which causes a random network to fall into pieces, only slightly shrinks the giant component of a scale-free network, while small fragments of the system become isolated. An attack, on the other hand, where the most connected nodes are the first to be removed, shows a different picture. While the Erdős–Rényi random network falls apart somewhat faster but in essentially the same way, a scale-free network is blasted apart by the removal of a much lower fraction of nodes. The results hold for simulations using the actual network topologies of the Internet and WWW [7,28].

The topological aspects of random node or edge removal can be calculated by mapping random failures to percolation problems. Cohen et al. have shown that uncorrelated random networks with a diverging second moment of their degree distribution (scale-free networks with

degree exponents between 2 and 3) have zero percolation thresholds [35]. Vázquez and Moreno used an approach that allowed them to investigate the percolation properties of correlated networks, showing that assortative mixing is beneficial for resilience: it can push the percolation threshold down to zero, even for networks with a finite second moment [144]. The opposite effect is also true: scale-free networks with diverging second moments but disassortative degree correlations are less resilient to node failure [89]. A general analytical approach based on generating function formalism not only reproduced the previous results, but also allowed Callaway and Newmann to investigate degree-dependent node removal scenarios, such as attacks [30]. Interestingly, assortative mixing does not help in case of an attack: Song et al. have found that non-fractal networks created via a mechanism that favors hub to hub connections on all scales are more vulnerable to intentional attack than their fractal (and also disassortative) counterparts [133].

**Cascading Failures** The function of a variety of complex networks involves the transmission or flow of some conserved quantity. Removal or failure of a node in such a network has consequences that ripple through the rest of the system far beyond the effects on connectivity: the node suddenly sheds the load or flux that it carried before, thus its neighbors suddenly experience higher loads: some of these overload and fail. A cascading failure can follow, as often seen on the power grid, perhaps the most quoted example of the phenomenon. Unlike the fragmentation of a network, a cascading failure can be triggered by a relatively small number of node failures, often a single one [73,100].

Using a model of overload in which the load-bearing capacity of a node is proportional to its load (betweenness) in the full network, Motter and Lai showed that scale-free networks are more vulnerable to randomly seeded cascading failures than random networks [100]. The vulnerability becomes especially pronounced under targeted attack: overload of the largest node can cause cascades that propagate through the whole system. They find that the most dangerous targets are not the most highly connected, but the most load-bearing (highest betweenness centrality) nodes. Betweenness is usually correlated with degree. The US powergrid, however, is vulnerable under attack targeting its most central nodes, but not its highest degree ones.

In a followup paper Motter introduced a fast defense strategy against cascading failures [99]. He argued that the only defense strategy which is fast enough is further removal of nodes: counterintuitive, but effective, if the nodes are carefully chosen. The reason this is possible, he argued,

is because nodes that carry small amounts of load (thus are less central to the network) actually generate much more load than they carry. Their shortest paths to the rest of the system are larger, thus all the communication (traffic, power) they receive or generate affects a larger number of intermediary nodes, increasing the total network load. Generated and handled load are anticorrelated, thus removal of nodes in ascending order of loads significantly reduces the size of cascading failures in the rest of the network, as verified by numerical simulations.

**Congestion** Cascading failures and congestion of traffic carried by networks are similar problems. As opposed to the node-removing effect of a failure, congestion does not disconnect a node, nonetheless, it can bring the traffic-bearing capability of a system to a halt. Most congestion models show a transition from free flow to congestion as the load is increased, regardless of the network type.

Ohira and Sawatari showed the occurrence of a jamming transition on a simple traffic model in which packets travel along the shortest paths of a two dimensional lattice between randomly chosen boundary nodes [113]. On a regular lattice shortest paths are highly degenerate, thus the authors considered two strategies: one is deterministic in picking its route, the other is random, preferring shortest paths but occasionally picking longer than optimal ones. They found that the jamming transition occurred at larger packet creation rates if the routing was probabilistic, with an optimal randomness that does not considerably lengthen the travel paths, but relieves the network from always choosing congested paths. Guimerá et al. considered a different formulation of the congestion problem on lattices and Cayley trees where all nodes generate packets to random destinations and send them along shortest paths, but the probability for a packet to hop between two nodes along the path depends on the queues accumulated by these two nodes [70]. Thus, there is some congestion-awareness built into the model. If the processing speed of nodes decreases with an increased number of packets, the system shows a discontinuous transition from free flow to congested state: the ratio of undelivered packets (the order parameter) jumps from 0 to 1. The positive feedback between congestion and slower processing leads to the formation of congestion nuclei that spread through the network. On the other hand, no transition is observed in networks in which the processing speed of nodes increases with queue lengths, only a crossover from low-density flow to high density flow accompanied by a change in fluctuation statistics. The critical case in between is queue-independent processing, which shows a continuous transition between free flow and congestion.

Solé and Valverde proposed a generalization of the Ohira–Sawatari model [131] and showed that the system at transition point exhibits self-similar time-series dynamics with an  $1/f$  power spectrum, as observed in measurements of packet transmission times on the Internet. Latency times and queue lengths also showed heavy tails with similar queue length distributions to jam size distribution in highway traffic models. Interestingly, the systems reaches its highest efficiency and information transfer regime right at the critical point, before entering the congested state [131].

Building on the observation that the time series dynamics of the model close to the critical point matches the real data, the authors introduced a self-organizing version of the model [142]. They argued that packet generation (thus user behavior) is linked to dynamics: each user tries to increase its rate until congestion of neighboring nodes is detected, at which point it starts to decrease it, dropping its rate to 0 if all its neighbors are congested. This model generates highly heterogeneous dynamics in space and time, with a power-law congestion length distribution. Moreover, this model points to internal network dynamics being responsible for fluctuations, supported by an analysis of real fluctuations by de Menezes and Barabási. They observed a power-law scaling of flux fluctuations as a function of total flux values, with two distinct scaling exponents.  $\alpha = 1/2$  scaling is generated by fluctuations internal to the system (as seen in the Internet and on a microchip), while  $\alpha = 1$  corresponds to external noise (seen in the WWW, river networks and the highway system) [43]. Solé and Valverde have further extended their model to study complex topologies similar to the Internet (using the model by Yook et al. [152]), with a routing strategy that can be tuned from random routing to global shortest path routing. Each node is assumed to know its neighborhood to  $m$  hops. If the destination of a packet is within a node's search horizon it uses shortest path routing, otherwise it passes the packet to a random neighbor [143] (local routing strategies with a fixed search horizon were also investigated by Tadić et al. [136,137]). They found that the routing was optimal when the search horizon equaled the average path length of the network. Further push toward global shortest paths actually decreased the efficiency by over-specifying paths and thus exacerbating congestion. The dynamics observed with optimal search depth reproduced the  $1/2$  scaling between fluctuations and mean flow observed in [43], as well as the power-law exponent of the average latency time distribution measured on the Internet.

Zhao et al. investigated the effect of network topology on congestion in a model in which the packet processing

speed of nodes is determined by their degree or their betweenness. They found that if the capacity of nodes was proportional to their degree, random networks as well as scale-free ones were less prone to congestion than regular lattices or Cayley trees. However, systems in which the node capacities were proportional to their betweenness had the same critical packet generation rate regardless of topology, suggesting that selectively increasing the capacity of high-betweenness nodes is a good way of increasing the carrying capacity of the system as a whole [154].

A different approach for investigating the effect of network topology on congestion was proposed by Toroczkai and Bassler [139]. They measured the congestion factor of the Erdős–Rényi and Barabási–Albert models, defined as the average fraction of nodes that do not receive and thus process incoming traffic. Instead of an explicit choice of routing or packet generation behavior, they assumed that packets on average follow the steepest gradients towards the neighbor with the highest “potential”, thus the gradient links determine the congestion factor (they assume a random distribution of node potentials). They found that the congestion factor of Erdős–Rényi random graphs increases with the size of the network, asymptotically growing to 1, while for scale-free networks it quickly reaches a value of  $\sim 0.7$  and does not increase with system size. A followup study by Danila et al. proposed a traffic model based on the idea of congestion-gradient driven flows [41].

The recognition that congestion occurs when the average number of packets processed by the busiest node reaches 1 (time-step) highlighted the importance of betweenness in the study of congestion. Routing protocols can influence the experienced betweenness of a node: the number of packets that actually go through it on average. Several methods of reducing the largest betweenness value have been investigated: optimization of link weights such that the shortest weighted paths give rise to a small maximum betweenness [40], hub avoiding global routing schemes [134] or traffic-aware routing [56]. Sreenivasan et al. proved that for any network topology there always exist an absolute upper bound for the communication threshold, determined by network topology, above which no routing algorithm can increase the critical congestion threshold [134].

## Spreading Processes and Social Dynamics

**Epidemic Models** Epidemiological modeling jumped to the forefront of networks research with a landmark paper by Pastor-Satorras and Vespignani which revealed the striking difference between virus spreading on scale-free

networks and homogeneous systems [115,116]. They studied the “susceptible–infected–susceptible” (SIS) epidemic model, suited to describe diseases that confer no immunity, such as tuberculosis, gonorrhea as well as computer viruses on systems that do not update their virus protection software. Each susceptible node can be infected with rate  $\nu$  if it is connected to one or more infected nodes, while infected nodes are cured with rate  $\delta$  and become susceptible again. The SIS model was known to show a non-equilibrium phase transition at a critical spreading rate  $\lambda_c$  ( $\lambda = \nu/\delta$ ): if the spreading rate is higher than this threshold, the disease is endemic; a finite fraction of nodes is persistently infected. Below the threshold the disease dies out completely after an initial breakout. As Pastor-Satorras and Vespignani pointed out, these results were obtained on lattices and failed to explain the very low but sustained prevalence of computer viruses. The authors report the absence of a critical point on scale-free networks with  $2 < \gamma \leq 3$ : there is no non-zero spreading rate for which the disease dies out. As  $\lambda$  approaches 0, the prevalence of the disease decreases exponentially, but only reaches 0 when  $\lambda = 0$ .

In a followup study together with Moreno they showed that the lack of an epidemic threshold holds for the “susceptible–infected–recovered” (SIR) model as well [96]. This model, applicable for diseases that result in immunity or death, has a radically different overall behavior from the SIS model. Recovered individuals cannot get infected again, thus the disease always dies out. Nonetheless, for homogeneous networks it shows a phase transition at a critical spreading rate  $\lambda_c$ , reflected in the total fraction of the population to ever get the disease. Below  $\lambda_c$  this fraction is vanishing as the system size goes to infinity, above  $\lambda_c$  there is a finite infected fraction. Moreno et al. have found that the critical spreading rate for uncorrelated complex networks is  $\lambda_c = \langle k \rangle / \langle k^2 \rangle$ .

Thus, networks with diverging connectivity fluctuations (such as scale-free graphs with  $2 < \gamma \leq 3$ ) have no non-zero epidemic threshold. Although real and thus finite-size networks cannot have infinite  $\langle k^2 \rangle$ , the epidemic threshold is very small for large systems, much smaller than that of a similar homogeneous network, and it decreases with systems size [117]. The SIR model has been mapped onto a bond percolation problem by Grassberger [68]. This mapping was exploited by Newman, who used the generating function formalism [30,112] to obtain the exact solution of the model in the infinite time limit for simple graphs with arbitrary degree distribution. He then extended the mapping to bipartite graphs to model sexually transmitted diseases spreading between men and women [106].

The absence of an epidemic threshold has profound implications on immunization policies effective in scale-free networks. Dezső and Barabási [47] as well as Pastor-Satorras and Vespignani [118] have shown that scale-free networks do not respond to random immunization: such a strategy cannot restore the epidemic threshold, or bring the effective spreading rate below it, even for an unrealistically high number of administered vaccines. On the other hand, immunization preferentially targeting the hubs of the network successfully reintroduces an epidemic threshold at very low vaccination rates, even if the strategy is imperfect in identifying the hubs. These results have important and controversial consequences for public policy of vaccination in the context of sexually transmitted diseases. This is partly because a policy that gives priority vaccination to prostitutes is controversial even if the public benefits are substantial, but also because it is generally difficult to identify the hubs of sexual networks. Cohen et al. proposed an immunization policy that overcomes this problem without relying on any global knowledge of the system: immunize a randomly chosen friend of a random person [36]. They showed that this strategy preferentially targets the hubs, thus restoring the epidemic threshold.

Epidemic spreading on networks with degree correlations show that degree correlations do not affect the lack of epidemic threshold in scale-free networks [23,97]. However, the epidemic incidence is smaller in these networks, while diseases are longer lived [97].

The influence of link weights on epidemic spreading has been investigated in the context of diseases such as Severe Acute Respiratory Syndrome (SARS), spreading on the world-wide airline network [37]. Colizza et al. developed a detailed multi-scale model of global epidemic spreading and found that the degree heterogeneity of the airport network accounts for most of the observed behavior of the epidemics, the heterogeneity of its weights has a much smaller influence on the spread of the disease [37]. They found that epidemics are fairly predictable, except at the very beginning of an outbreak originating from a hub, as well as the end of epidemics [37].

**Information Spreading** The similarity between rumor spreading and epidemic spreading was recognized as early as 1964, when Goffman and Newill used the SIR model to describe the spread of ideas instead of diseases [65]. The authors pointed out an important difference between the two processes: while one studies epidemic models with an effective immunization strategy in mind, information spread is in general favored; moreover, the process can be engineered in a way that facilitates high reliability of the spread. Later in the same year, Daley and Kendall pub-

lished a spreading model that is specific to rumor (or information) spreading: instead of the random recovery rate of the SIR model (which assumes spontaneous forgetting of the news), they assumed that nodes lose their interest in spreading the news upon encounter with another node that knows it [38]. Thus the “recovery” process of the Daley–Kendall model is very different than that of SIR, leading to a significant difference in the outcome of spreading. Assuming homogeneous mixing (each infected individual, or spreader can randomly contact any node in the system), the final fraction of nodes who have heard the news is  $\sim 80\%$ .

A decentralized, spontaneous and robust method of spreading information was welcome by the computer science community; it allowed them to overcome the scalability problems of data disseminating protocols in large-scale distributed computing [146] and peer-to-peer networks [78]. It was also used in update management of databases duplicated at many sites [46].

In the context of complex networks, the Daley–Kendall model was first investigated on the small-world model by Zanette, who found that below a density of shortcuts the rumor dies out without reaching a finite fraction of nodes [153]. Moreover, this threshold is finite for infinite systems ( $p_c = 0.2$  for  $K = 2$ ), even though the network is turned into a small-world by a vanishingly small number of shortcuts. Moreno et al. studied the same model on homogeneous versus scale-free networks and found that as opposed to epidemic spreading, rumors spread to a higher fraction of nodes in homogeneous networks [98]. Hubs in scale-free networks have a conflicting effect on rumor spreading: while they are highly likely and quick to hear the information, they also facilitate the formation of stifler nodes who do not spread the rumor.

## Searching on Complex Networks

Milgram’s famous experiment showing six degrees of separation between two random people not only showed the existence of short paths on social networks, it also highlighted the fact that finding a destination node with local information is not very difficult [94]. Motivated by these findings, Kleinberg showed that the performance of a simple greedy search critically depends on the topology of the underlying network. He used a small-world type model in which nodes sit on a two-dimensional lattice (connected to their four neighbors) to which a few long-range connections are added. Unlike in the small-world model, these connections are not added entirely randomly: the probability of a shortcut is proportional to  $r^{-\alpha}$ , where  $r$  is the Manhattan distance between two nodes. In this model the

best-case performance of a decentralized algorithm (the expected delivery time) scales as  $(\log N)^2$  for  $\alpha = 2$ . However, both larger and smaller values of  $\alpha$  result in expected times that scale as polynomials in  $N$ , highlighting that not all networks with short average path lengths allow the use of efficient decentralized searches [81].

Adamic et al. proposed to exploit the heterogeneous nature of scale-free networks to significantly speed up breath-first type searches (also called burning algorithms) which typically run in  $\mathcal{O}(N)$  time [3,80]. Instead of broadcasting a query to all neighbors who in turn broadcast it further, their algorithm checks whether the target is among its first neighbor, and if not, it passes the query along to the neighbor with the highest degree. This neighbor continues the search in a similar way, and if a dead end is reached, the message is recursively back-tracked until there are no neighbors left to explore. This strategy can improve the performance to  $\mathcal{O}(N^{1/2})$  for  $\gamma = 3$ , and  $\mathcal{O}(\log N)$  for  $\gamma = 2$  scale-free networks. The method was tested by the authors on the GNUTELLA peer-to-peer network, as well as by Kim et al. on scale-free networks generated by the configuration model and the Barabási-Albert model [80]. (Kim, Yoon, Han and Jeong proposed the strategy independently a few months after Adamic et al. submitted their paper.)

Searchability of social networks has been revisited several times since Milgram's experiment. Killworth and Bernard found that the choice of neighbor to pass the letter on was most often based on common characteristics with the target, such as geographical location and profession [79]. An electronic version of the experiment carried out by Dodds et al. concluded that the search on the e-mail network is also guided by reliance on common "social identities", and the chosen routes do not favor social hubs [49]. Motivated by these results, Watts et al. and Kleingerg independently proposed a searchable social network model in which nodes are grouped in a nested hierarchy of social categories. A link between two nodes is exponentially less likely if their social distance (length of the path along the hierarchical tree of social groups) is large [82,149]. In a simulated Milgram-type experiment, the letter would be passed from a node to a neighbor with the smallest social distance to the target. Such a search performs well for a range of the model parameters, with its best performance scaling as  $\mathcal{O}(\log N)$  [82].

### Future Directions

The short account of research advances into dynamics on complex networks presented here is by no means exhaustive. Several fairly large and lively research areas, synchro-

nization, boolean networks, random walks on complex networks, brain networks, models of adaptive or dynamical wiring among them, have all been omitted here.

Future directions in networks research are hard to account for in a few paragraphs. It is believed that further understanding of dynamics on complex networks is the general direction of the field. Along with a shift from pure structural studies to dynamics, there has also been a shift from studies of networks in general and features that are common to most of them to more application-driven studies of increasingly narrow classes of networks. This is not to say that important lessons learned this way do not often carry over from one system to the other, but it shows that enough is known about network characteristics now to tell them apart, to look for distinguishing features as well as universal ones.

A fairly standard toolkit for the assessment of the structural properties of a network has already emerged, making the notion of networks and knowledge about them a useful toolkit in studying the architecture of complex systems. Our hope is that dynamical studies will further extend the toolkit to cover general aspects of the type of events occurring on complex networks, and perhaps lead the way in understanding how function emerges in complex systems.

## Bibliography

### Primary Literature

1. Adamic LA (1999) The small world web. In: Lecture Notes in Computer Science vol 1696. Springer, New York, pp 443–454
2. Adamic LA, Huberman BA (2000) Power-law distribution of the World Wide Web. *Science* 287:2115
3. Adamic LA, Lukose RM, Puniyani AR, Huberman BA (2001) Search in power-law networks. *Phys Rev E* 64(4):046 135
4. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. In: Proceedings of the 32nd Annual ACM Symposium on Theory of Computing. ACM, New York, pp 171–180
5. Albert R, Barabási A-L (2000) Topology of evolving networks: local events and universality. *Phys Rev Lett* 85:5234
6. Albert R, Jeong H, Barabási A-L (1999) Diameter of the World-Wide Web. *Nature* 401:130–131
7. Albert R, Jeong H, Barabási A-L (2000) Attack and error tolerance of complex networks. *Nature* 406:378
8. Albert R, Albert I, Nakarado GL (2004) Structural vulnerability of the North American power grid. *Phys Rev Lett* 69:025 103
9. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási A-L (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 427(6977):839–843
10. Alon U (2007) Network motifs: theory and experimental approaches. *Nat Rev Genet* 8:450–461
11. Amaral LAN, Scala A, Barthélemy M, Stanley HE (2000) Classes of small-world networks. *Proc Natl Acad Sci USA* 97:11 149



12. Arenas A, Danon L, Díaz-Guilera A, Gleiser PM, Guimerà R (2004) Community analysis in social networks. *Euro Phys J B* 38(2):373–380
13. Baiesi M, Paczuski M (2004) Scale-free networks of earthquakes and aftershocks. *Phys Rev E* 69(6):066106
14. Balázi G, Barabási A-L, Oltvai ZN (2005) Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc Natl Acad Sci USA* 102(22):7841–7846
15. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
16. Barabási A-L, Albert R, Jeong H (1999) Mean-field theory for scale-free random networks. *Physica A* 272:173–187
17. Barabási A-L, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. *Physica A* 311:590
18. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci USA* 101:3747–3752
19. Barthélemy M (2003) Crossover from scale-free to spatial networks. *EuroPhys Lett* 63(6):915–921
20. Bender EA, R Canfield E, McKay BD (1997) The asymptotic number of labeled graphs with  $n$  vertices,  $q$  edges, and no isolated vertices. *J Comb Theor: Series A* 80(1):124–150
21. Bianconi G, Barabási A-L (2001) Competition and multiscaling in evolving networks. *EuroPhys Lett* 54:436
22. Bianconi G, Barabási A-L (2001) Bose-Einstein condensation in complex networks. *Phys Rev Lett* 86:5632
23. Boguñá M, Pastor-Satorras R, Vespignani A (2003) Absence of epidemic threshold in scale-free networks with degree correlations. *Phys Rev Lett* 90(2):028 701
24. Boguñá M, Pastor-Satorras R, Diaz-Guilera A, Arenas A (2004) Models of social networks based on social distance attachment. *Phys Rev E* 70(5):056122
25. Bollobás B, Riordan O (2004) The diameter of a scale-free random graph. *Combinatorica* 24(1):5–34
26. Bollobás B, de la Vega WF (1982) The diameter of random regular graphs. *Combinatorica* 2(2):125–134
27. Bollobás B, Riordan O, Spencer J, Tusnády G (2001) The degree sequence of a scale-free random process. *Random Struct Algorithms* 18:279–290
28. Broder A, Kumar R, Maghoul F, Raghavan P, Rajalopagan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the web. *Comput Netw* 33:309–320
29. Broida A, Claffy KC (2001) Internet topology: Connectivity of IP graphs. In: Fahmy S, Park K (eds) *Scalability and Traffic Control in IP Networks in Proc SPIE*, vol 4526. International Society for Optical Engineering, Bellingham, pp 172ñ–187
30. Callaway DS, Newman MEJ, Strogatz SH, Watts DJ (2000) Network robustness and fragility: Percolation on random graphs. *Phys Rev Lett* 85:5468
31. Capocci A, Servedio VDP, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G (2006) Preferential attachment in the growth of social networks: the internet encyclopedia wikipedia. *Phys Rev E* 74:036 116
32. Chung F, Lu L (2002) The average distances in random graphs with given expected degrees. *Proc Natl Acad Sci USA* 99:15879–15882
33. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. *Phys Rev E* 70(6 Pt 2):066 111
34. Cohen R, Havlin S (2003) Scale-free networks are ultra small. *Phys Rev Lett* 90:058 701
35. Cohen R, Erez K, ben Avraham D, Havlin S (2000) Resilience of the Internet to random breakdowns. *Phys Rev Lett* 85:4626–4628
36. Cohen R, Havlin S, ben Avraham D (2003) Efficient immunization strategies for computer networks and populations. *Phys Rev Lett* 91(24):247 901
37. Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci USA* 103(7):2015–2020
38. Daley DJ, Kendall DG (1965) Stochastic rumours. *IMA J Appl Math* 1(1):42–55
39. Dall J, Christensen M (2002) Random geometric graphs. *Phys Rev E* 66:016 121
40. Danila B, Yu Y, Marsh JA, Bassler KE (2006) Optimal transport on complex networks. *Phys Rev E* 74(4):046106
41. Danila B, Yu Y, Marsh JA, Bassler KE (2007) Transport optimization on complex networks. *Chaos* 17(2):026102
42. Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a small world from local interactions: Modeling acquaintance networks. *Phys Rev Lett* 88(12):128 701
43. de Menezes MA, Barabási A-L (2004) Fluctuations in network dynamics. *Phys Rev Lett* 92:028 701
44. de Solla Price DJ (1965) Networks of scientific papers. *Science* 149:510–515
45. de Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. *J Am Soc Inform Sci* 27:292–306
46. Demers A, Greene D, Hauser C, Irish W, Larson J, Shenker S, Sturgis H, Swinehart D, Terry D (1987) Epidemic algorithms for replicated database maintenance. In: *PODC '87: Proc. 6th Ann. ACM Symposium on Principles of distributed computing*. ACM, New York, pp 1–12
47. Dezső Z, Barabási A-L (2002) Halting viruses in scale-free networks. *Phys Rev E* 65:055 103
48. Di Matteo T, Aste T, Gallegati M (2005) Innovation flow through social networks: productivity distribution in france and italy. *Euro Phys J B* 47(3):459–466
49. Dodds PS, Muhamad R, Watts DJ (2003) An experimental study of search in global social networks. *Science* 301(5634):827–829
50. Dorogovtsev SN, Mendes JFF (2000) Scaling behaviour of developing and decaying networks. *Europhys Lett* 52:33
51. Dorogovtsev SN, Mendes JFF (2001) Effect of the accelerating growth of communications networks on their structure. *Phys Rev E* 63:025 101
52. Dorogovtsev SN, Mendes JFF (2001) Language as an evolving word web. *Proc R Soc London B* 268:2603–2606
53. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Structure of growing networks with preferential linking. *Phys Rev Lett* 85:4633–4636
54. Dorogovtsev SN, Goltsev AV, Mendes JFF (2002) Pseudofractal scale-free web. *Phys Rev E* 65:066 122
55. Ebel H, Mielsch LI, Bormholdt S (2002) Scale-free topology of e-mail networks. *Phys Rev E* 66:035 103
56. Echenique P, Gómez-Gardeñes J, Moreno Y (2004) Improved routing strategies for internet traffic delivery. *Phys Rev E* 70(5):056105

57. Erdős P, Rényi A (1959) On random graphs I. *Publ Math (Debrecen)* 6:290–297
58. Erdős P, Rényi A (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* 5:17–61
59. Euler L (1741) *Solutio problematis ad geometriam situs pertinentis*. *Commentarii academiae scientiarum Petropolitanae* 8:128–140
60. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the Internet topology. *Comput Commun Rev* 29:251–262
61. Ferrer i Cancho R, Solé RV (2001) The small-world of human language. *Proc R Soc London B* 268:2261–2266
62. Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry* 40(1):35–41
63. Gerdes SY, Scholle MD, Campbell JW, Balázs G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatal V, DiSouza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási A-L, Oltvai ZN, Osterman AL (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J Bacteriol* 185:5673–5684
64. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821–7826
65. Goffman W, Newill VA (1964) Generalization of epidemic theory: An application to the transmission of ideas. *Nature* 204(4955):225–228
66. Goh K-I, Oh E, Jeong H, Kahng B, Kim D (2002) Classification of scale-free networks. *Proc Natl Acad Sci USA* 99(20):12583–12588
67. Govindan R, Tangmunarunkit H (2000) Heuristics for Internet map discovery. In: *Proceedings of IEEE INFOCOM 2000*. Nineteenth annual joint conference of the IEEE computer and communications societies, Tel Aviv, Israel, vol 3. IEEE, Piscataway, New Jersey, pp 1371–1380
68. Grassberger P (1983) Critical behavior of the general epidemic process and dynamical percolation. *Math Biosci* 63(2):157–172
69. Guimerà R, Amaral LAN (2004) Modeling the world-wide airport network. *Euro Phys J B* 38(2):381–385
70. Guimerà R, Arenas A, Díaz-Guilera A, Giralt F (2002) Dynamical properties of model communication networks. *Phys Rev E* 66(2):026704
71. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. *Phys Rev E* 68(6):065103
72. Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. *Phys Rev E* 65(2):026107
73. Holme P, Kim BJ (2002) Vertex overload breakdown in evolving networks. *Phys Rev E* 65(6):066109
74. Jain S, Krishna S (1998) Autocatalytic sets and the growth of complexity in an evolutionary model. *Phys Rev Lett* 81(25):5684–5687
75. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L (2000) The large-scale organization of metabolic networks. *Nature* 407:651–654
76. Jeong H, Néda Z, Barabási A-L (2003) Measuring preferential attachment for evolving networks. *EuroPhys Lett* 61:567
77. Jung S, Kim S, Kahng B (2002) A geometric fractal growth model for scale free networks. *Phys Rev E* 65:056101
78. Kermarrec AM, Massoulie L, Ganesh AJ (2003) Probabilistic reliable dissemination in large-scale systems. *IEEE Trans Parallel Distributed Syst* 14(3):248–258
79. Killworth PD, Bernard HR (1978) The reverse small world experiment. *Social Netw* 1:159–192
80. Kim BJ, Yoon CN, Han SK, Jeong H (2002) Path finding strategies in scale-free networks. *Phys Rev E* 65(2):027103
81. Kleinberg JM (2000) Navigation in a small world. *Nature* 406(6798):845
82. Kleinberg JM (2002) Small-world phenomena and the dynamics of information. In: Dietterich TG, Becker S, Ghahramani Z (eds) *Proceedings of the 2001 Neural Information Processing Systems Conference*. MIT Press, Cambridge
83. Kleinberg JM, Kumar SR, Raghavan P, Rajagopalan S, Tomkins A (1999) The web as a graph: Measurements, models and methods. In: *Proc. of the Int. Conf. on Combinatorics and Computing, COCOON'99 Berlin*. Springer, Tokyo, p 1
84. Klemm K, Eguíluz VM (2002) Growing scale-free networks with small-world behavior. *Phys Rev E* 65:057102
85. Krapivsky PL, Redner S (2001) Organization of growing random networks. *Phys Rev E* 63:66–123
86. Krapivsky PL, Redner S (2002) A statistical physics perspective on web growth. *Comput Netw* 39:261–276
87. Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. *Phys Rev Lett* 85:4629–4632
88. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Trawling the web for emerging cyber-communities. *Comput Netw* 31:1481–1493
89. Leone M, Vázquez A, Vespignani A, Zecchina R (2002) Ferromagnetic ordering in graphs with arbitrary degree distribution. *Euro Phys J B* 28:191–197
90. Liljeros F, Edling C, Amaral L, Aberg Y (2001) The web of human sexual contacts. *Nature* 411:907–908
91. Ma HW, Zeng AP (2003) The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 19(11):1423–1430
92. Manna SS, Sen P (2002) Modulated scale-free network in euclidean space. *Phys Rev E* 66(6):066114
93. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296:910–913
94. Milgram S (1967) The small-world problem. *Psychology Today* 2:60–67
95. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. *Science* 298:824–827
96. Moreno Y, Pastor-Satorras R, Vespignani A (2002) Epidemic outbreaks in complex heterogeneous networks. *Euro Phys J B* 26(4):521–529
97. Moreno Y, Gómez JB, Pacheco AF (2003) Epidemic incidence in correlated complex networks. *Phys Rev E* 68(3):035103
98. Moreno Y, Nekovee M, Vespignani A (2004) Efficiency and reliability of epidemic data dissemination in complex networks. *Phys Rev E* 69(5):055101
99. Motter AE (2004) Cascade control and defense in complex networks. *Phys Rev Lett* 93(9):098701
100. Motter AE, Lai YC (2002) Cascade-based attacks on complex networks. *Phys Rev E* 66(6):065102
101. Motter AE, de Moura APS, Lai YC, Dasgupta P (2002) Topology of the conceptual network of language. *Phys Rev E* 65:065102

102. Myers CR (2003) Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs. *Phys Rev E* 68:046116
103. Newman MEJ (2001) Clustering and preferential attachment in growing networks. *Phys Rev E* 64:025102(R)
104. Newman MEJ (2001) The structure of scientific collaboration networks. *Proc Natl Acad Sci USA* 98:404–409
105. Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett* 89:208701
106. Newman MEJ (2002) Spread of epidemic disease on networks. *Phys Rev E* 66(1):016128
107. Newman MEJ (2003) Mixing patterns in networks. *Phys Rev E* 67:026126
108. Newman MEJ (2004) Fast algorithm for detecting community structure in networks. *Phys Rev E* 69(6 Pt 2):066133
109. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 74(3):036104
110. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69(2):026113
111. Newman MEJ, Watts DJ (1999) Renormalization group analysis of the small-world network model. *Phys Lett A* 263:341–346
112. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Phys Rev E* 64(2):026118
113. Ohira T, Sawatari R (1998) Phase transition in a computer network traffic model. *Phys Rev E* 58(1):193–195
114. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435(7043):814–818
115. Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. *Phys Rev E* 63(6):066117
116. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Phys Rev Lett* 86:3200–3203
117. Pastor-Satorras R, Vespignani A (2002) Epidemic dynamics in finite size scale-free networks. *Phys Rev E* 65(3):035108
118. Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. *Phys Rev Lett* 65:036104
119. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. *Phys Rev Lett* 87:258701
120. Rao F, Caflisch A (2004) The protein folding network. *J Mol Biol* 342:299–306
121. Ravasz E, Barabási A-L (2002) Hierarchical organization in complex networks. *Phys Rev E* 67:026122
122. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297:1551–1555
123. Ravasz E, Gnanakaran S, Toroczkai Z (2007) Network structure of protein folding pathways. *arXiv:0705.0912v1*
124. Redner S (1998) How popular is your paper? An empirical study of the citation distribution. *Euro Phys J B* 4:131–135
125. Redner S (2004) Citation statistics from more than a century of Physical Review. *arXiv:physics/0407137v2*
126. Scala A, Amaral LAN, Barthélémy M (2001) Small-world networks and the conformation space of a short lattice polymer chain. *Europhys Lett* 55(4):594–600
127. Serrano MA, Boguñá M (2003) Topology of the world trade web. *Phys Rev E* 68:015101
128. Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *E. coli*. *Nat Genet* 31:64–68
129. Sigman M, Cecchi GA (2002) Global organization of the Wordnet lexicon. *Proc Natl Acad Sci USA* 99(3):1742–1747
130. Solé R, Pastor-Satorras R, Smith E, Kepler T (2002) A model of large-scale proteome evolution. *Adv Compl Syst* 5:43–54
131. Solé RV, Valverde S (2001) Information transfer and phase transitions in a model of internet traffic. *Physica A* 289(3–4):595–605
132. Song C, Havlin S, Makse HA (2005) Self-similarity of complex networks. *Nature* 433(7024):392–395
133. Song C, Havlin S, Makse HA (2006) Origins of fractality in the growth of complex networks. *Nat Phys* 2(4):275–281
134. Sreenivasan S, Cohen R, Lopez E, Toroczkai Z, Stanley HE (2007) Structural bottlenecks for communication in networks. *Phys Rev E* 75(3):036105
135. Tadić B (2001) Dynamics of directed graphs: The world-wide web. *Physica A* 293:273–284
136. Tadić B, Rodgers GJ (2002) Packet Transport on Scale Free Networks. *Adv Compl Syst* 5:445–456
137. Tadić B, Thurner S (2004) Information super-diffusion on structured networks. *Physica A* 332:566–584
138. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* 20(5):433–440
139. Toroczkai Z, Bassler KE (2004) Network dynamics: Jamming is limited in scale-free systems. *Nature* 428:716
140. Toroczkai Z, Kozma B, Bassler KE, Hengartner NW, Korniss G (2004) Gradient networks. *arXiv:cond-mat/0408262v1*
141. Tyler JR, Wilkinson DM, Huberman BA (2003) Email as spectroscopy: Automated discovery of community structure within organizations. In: Huysman M, Wenger E, Wulf V (eds) *Communities and Technologies. Proceedings of the First International Conference on Communities and Technologies*. Kluwer, Norwell MA, pp 81–96
142. Valverde S, Solé RV (2002) Self-organized critical traffic in parallel computer networks. *Physica A* 312(3–4):636–648
143. Valverde S, Solé RV (2004) Internet's critical path horizon. *Euro Phys J B* 38(2):245–252
144. Vázquez A, Weigt M (2003) Computational complexity arising from degree correlations in networks. *Phys Rev E* 67(2):027101
145. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modelling of protein interaction networks. *ComplexUs* 1:38–44
146. Vogels W, van Renesse R, Birman K (2003) The power of epidemics: robust communication for large-scale distributed systems. *SIGCOMM Comput Commun Rev* 33(1):131–135
147. Wasserman S, Faust K (1994) *Social Network Analysis*. Cambridge University Press, Cambridge
148. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. *Nature* 393:440–442
149. Watts DJ, Dodds PS, Newman MEJ (2002) Identity and search in social networks. *Science* 296:130
150. Wilkinson DM, Huberman BA (2004) A method for finding communities of related genes. *Proc Natl Acad Sci USA Suppl* 101(1):5241–5248
151. Xulvi-Brunet R, Sokolov IM (2002) Evolving networks with disadvantaged long-range connections. *Phys Rev E* 66(2):026118

152. Yook SH, Jeong H, Barabási A-L (2003) Modelling the Internet's large-scale topology. *Proc Natl Acad Sci USA* 99:13382–13386
153. Zanette DH (2001) Critical behavior of propagation on small-world networks. *Phys Rev E* 64(5):050901
154. Zhao L, Lai YC, Park K, Ye N (2005) Onset of traffic congestion in complex networks. *Phys Rev E* 71(2):026125

### Books and Reviews

- Barabási A-L (2002) *Linked: The New Science of Networks*. Perseus Publishing, Cambridge
- Pastor-Satorras R, Rubi M, Diaz-Guilera A (eds) (2003) *Statistical Mechanics of Complex Networks*, 625, *Lecture Notes in Physics*. Springer, Berlin
- Mendes J, Oliveira JG, Abreu FV, Povolotsky A, Dorogovtsev SN (eds) (2005) *Science of Complex Networks: From Biology to the Internet and WWW*. AIP conference proceedings, CNET 2004, Aveiro, Portugal, vol 776
- Newman MEJ, Barabási A-L, Watts DJ (eds) (2003) *The Structure and Dynamics of Complex Networks*. Princeton University Press, Princeton
- Ben-Naim E, Frauenfelder H, Toroczkai Z (eds) (2005) *Complex Networks*, 650, *Lecture Notes in Physics*. Springer, Secaucus
- Bornholdt S, Schuster HG (eds) (2002) *Handbook of graphs and networks: from the genome to the internet*. Wiley-VCH, Berlin
- Bollobás B (1985) *Random Graphs*. Academic Press, London
- Watts DJ (1999) *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton
- Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74(1):47–97
- Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45(2):167–256
- Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. *Adv Phys* 51:1079
- Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. *Phys Rep* 424(4–5):175–308

## Neuro-fuzzy Control of Autonomous Robotics

PETRU EMANUEL STINGU, FRANK L. LEWIS  
Automation & Robotics Research Institute,  
University of Texas at Arlington, Fort Worth, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Control Hierarchy  
Map-Based Navigation  
Behaviors  
Future Directions  
Bibliography

### Glossary

**Robot** The word ‘robot’ was introduced by the Czech playwright Capek in his 1920 play *Rossum’s Universal Robots*. The word ‘robota’ in Czech means simply ‘work’. Although there is no definition accepted by everyone, in this chapter a robot is considered to be a human-built machine that is mobile, can sense and interact with the environment, and has the necessary intelligence in order to handle unforeseen circumstances autonomously. Most important than all, it has to do a useful task.

**Autonomy** Independence of control, self-sufficiency. Applied to robots, it implies the ability of the robot to find solutions by itself to the various problems that might appear while completing the assigned task.

**Fuzzy logic** The idea of fuzzy logic was first advanced by Dr. Lotfi Zadeh of the University of California at Berkeley in the 1960s. It came from the fact that natural language can not be easily translated in the absolute terms of 0 and 1. Fuzzy logic includes 0 and 1 as extreme cases of truth (that are representations of certainty or facts), but also includes the various states of truth in between (partial truth). As an example, using binary logic it can be said that “the target is on the left side of the robot” or “the target is not on the left side of the robot”, while using fuzzy logic a more precise description can be given, like “the target is 20% on the left side of the robot”.

**Neural networks** A neural network can be described as a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes. Some of its advantages are: non-linear mappings, adaptation and learning, ease of implementation and self-organization.

**Control** When talking about controlling systems, control refers to the process of changing the input of the system such that its output reaches a desired value. Most of the time, the control is done in a closed loop, where the output of the system is continuously compared to the reference and the necessary control command is applied to the inputs of the system in order to reduce the error at the output.

**Behavior** Small independent decision-making process that fully implements a control policy for one specific sub-task. Usually multiple behaviors coexist and are enabled or disabled by an arbiter, depending on which is useful in a particular situation.



**Mapping** Mapping is split into two main processes that are dependent on each other: map-learning and localization. The first has to store the information from the robot sensors into a suitable internal representation (map). The latter has to estimate the position of the various objects on the map. Building the map needs localization, but in the same time localization requires a map.

**Path planning** The process where, given a complete description of the geometry of a robot and the static environment populated with obstacles, a collision-free path must be found such that the robot can move from an initial position and orientation to a goal position and orientation.

### Definition of the Subject

Autonomous robots are robots which can perform desired tasks in unstructured environments without requiring continuous human guidance. Most of the times, the dynamics of the robot itself can be described analytically. Unfortunately, in many robotic applications, it is difficult if not impossible to obtain a precise mathematical model of the environment and its interaction with the robot through actuators and sensors. The lack of complete and precise knowledge about the environment limits the applicability of conventional control system design to the domain of autonomous robotics. Some of the requirements for a robot to successfully achieve autonomy are the possibility to acquire knowledge about the environment and itself, to reason under uncertainty and to have learning capabilities in order to adapt to the environment based on accumulated experience.

Efficient control algorithms for autonomous robots should imitate the way humans are operating manned or similar vehicles. When making decisions, humans tend to work with vague or imprecise concepts that can often be expressed linguistically. Lotfi Zadeh has proposed one way to model this decision making process by introducing the fuzzy set theory in the field of control [36]. Fuzzy logic is particularly suited for problems in which the data, the objectives and the constraints are too complex or too ill-defined to admit a precise mathematical analysis.

Neural networks were developed as an attempt to realize mathematical models of brain-like systems. The key advantage is their ability to learn from examples instead of requiring an algorithmic development from the designer. They can generalize to new situations. Once a neural network has been trained for a set of data, it can interpolate and produce answers for the cases not present in the training set.

While fuzzy logic can be used to represent knowledge in a human-readable form and to use it for reasoning, neural networks allow adaptation and learning in dynamic environments under varying conditions. Neuro-fuzzy techniques combine the advantages of both methods by having neural networks adapt the knowledge base of the fuzzy logic systems or fuzzy systems tune the weights of neural networks. These relatively simple control methods can be used to successfully implement complex intelligent autonomous robots, robust to uncertainties in their own model, in the environment and in the readings from the sensors.

### Introduction

Despite the differences between neural networks and fuzzy logic systems, they can actually be unified at the level of the universal function approximator. Both of them define a nonlinear function  $y = f(x)$  from inputs to outputs and if designed properly satisfy the “universal approximation property” [22,32]. Then, for any continuous function  $\psi(x)$  defined on a closed and bounded set and an arbitrary number  $\varepsilon > 0$ , there exists a neural network or a fuzzy system  $f(x)$  such that

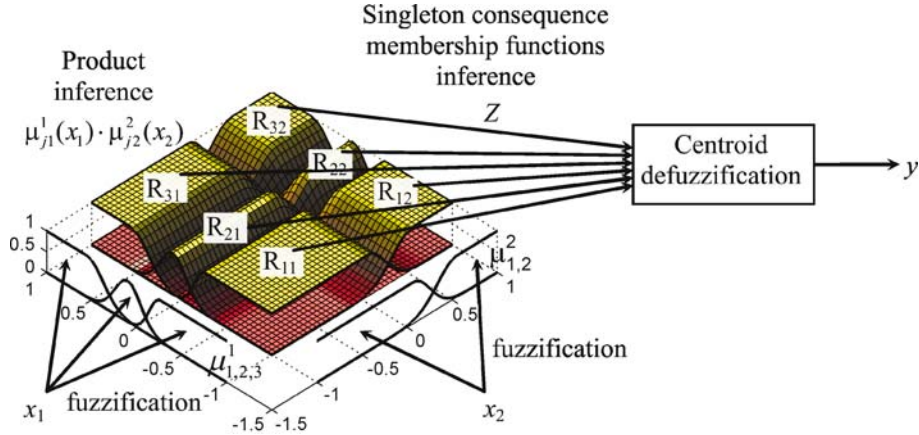
$$\sup_x |f(x) - \psi(x)| < \varepsilon$$

This property does not say how to build the neural net or the fuzzy system. It simply shows that with the proper structure and with enough tuning it is possible to approximate a continuous function with an error that can be made as small as desired.

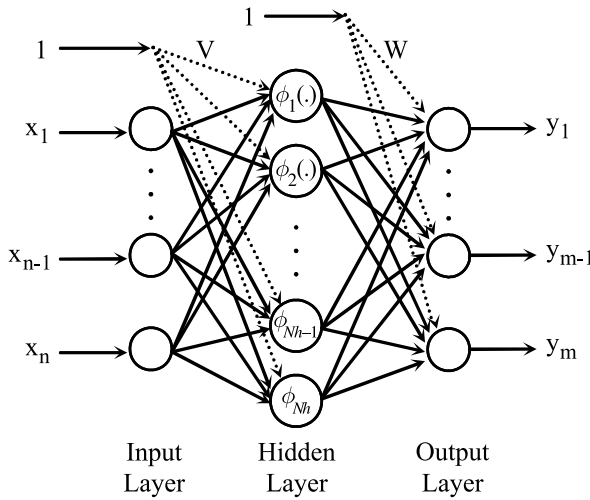
The multilayer perceptron should be viewed as a non-linear network whose non-linearity can be tuned by changing the weights, biases and the parameters of the activation functions. The fuzzy logic system is also a tunable non-linearity whose shape can be tuned by changing the membership functions. In both cases, it is possible to use gradient methods for tuning. The back-propagation training method [33] is well-known from the neural networks field and inspired gradient training of fuzzy systems.

Some radial basis function neural networks are functionally equivalent to some standard fuzzy systems in the sense that given the same inputs, they will produce the same outputs. This can be shown on a normalized RBF neural net that has the same number of neurons on the hidden layer as the number of rules in the fuzzy system with product inference and centroid defuzzification. A simple example for a two-input fuzzy system is shown in Fig. 1. A RBF neural net structure that can process the information in a similar way is shown in Fig. 2.





Neuro-fuzzy Control of Autonomous Robotics, Figure 1  
Example of a two-variable fuzzy logic system



Neuro-fuzzy Control of Autonomous Robotics, Figure 2  
Multi-layer neural network

The output of a fuzzy system with  $n$  input variables, possible different number of membership functions for each input variable, product inference, control representative values  $z_{i,j1,j2,\dots,jn}$  and centroid defuzzification is the following:

$$y = \frac{\sum_{j1,j2,\dots,jn} z_{i,j1,j2,\dots,jn} \prod_{k=1}^n \mu_{jk}^k(x_k)}{\sum_{j1,j2,\dots,jn} \prod_{k=1}^n \mu_{jk}^k} \quad (1)$$

The  $k$ th output of an unnormalized RBF neural network with  $N_h$  neurons on the hidden layer is

$$y_k = \sum_{i=1}^n w_{ik} \cdot \phi_i(\mathbf{x}) \quad (2)$$

while the output of a multi-layer neural network is

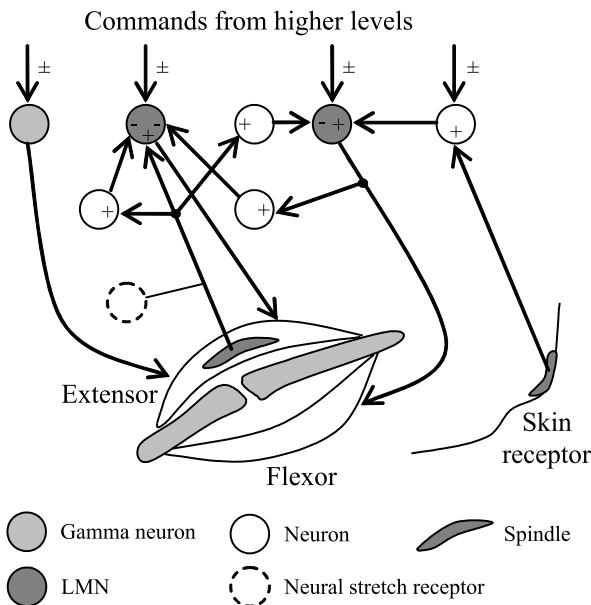
$$\mathbf{y} = \mathbf{W}^T \boldsymbol{\phi}(\mathbf{V}^T \mathbf{x}). \quad (3)$$

Because of their properties of learning, recall, function approximation, generalization, classification, association, pattern recognition and clustering, neural networks and fuzzy logic systems can be used to solve a large set of problems. They are successfully used in all types of autonomous robots. Some examples are:

- ground robots: indoor and outdoor
- aquatic robots: surface and submersible
- aerial robots: fixed-wing and rotary-wing

### Control Hierarchy

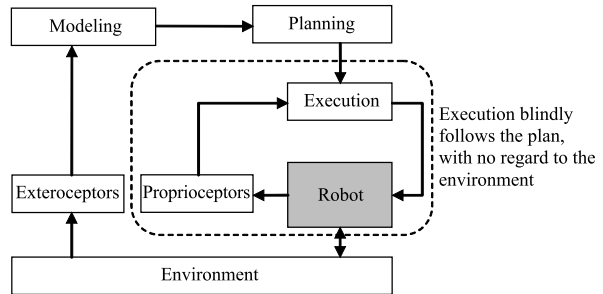
We have seen that autonomous robots can borrow traits from humans. Fuzzy logic copies the way they represent knowledge and the mechanisms of logical reason. Neural networks implement a simple model for the low-level organization and the physiological mechanisms relating to information processing in the nervous system. It is a well-known fact that nature has found the optimal solutions for every imaginable problem, over thousands and thousands of years of refinements. Humans have become the most autonomous of all living beings. They can navigate all around the Earth and even in the extraterrestrial space, they gather knowledge and use it to get even further. Such a successful model should also suggest the control hierarchy to be used on autonomous robots. The organization of the human nervous system involved in movement can provide invaluable information on how to design an efficient hierarchical control architecture.



**Neuro-fuzzy Control of Autonomous Robotics, Figure 3**  
Joint with the principal neurons involved in reflex control

Consideration of the vast and complex system of structures and pathways involved in movement will begin at its lowest end, the spinal cord, where more basic motor control is localized. The most important human interoceptive reflex is the myotatic reflex [4], which originates from the neuro-muscular fibers. Its principal function is maintaining the joint position fixed and compensating external noise. In Fig. 3, suppose that a load is applied to the joint. This will flex the joint, causing the stretching of the extensor muscle and also stretching of the spindle. This will increase the output of the neural stretch receptor neuron, which increases the output of the local motor neuron (LMN). The resulting increase in the contraction force will compensate the load. This local feedback allows the higher system to ignore the fluctuation in contraction required to maintain a certain joint extension. On a robot, the similar function is done by the low-level controllers for the motors. For example, they receive a reference speed and they have to keep the output speed equal with the reference, independent of the torque disturbances. Neural nets and fuzzy logic speed regulators have been successfully applied for motor control, yielding better results than normally used PID controllers.

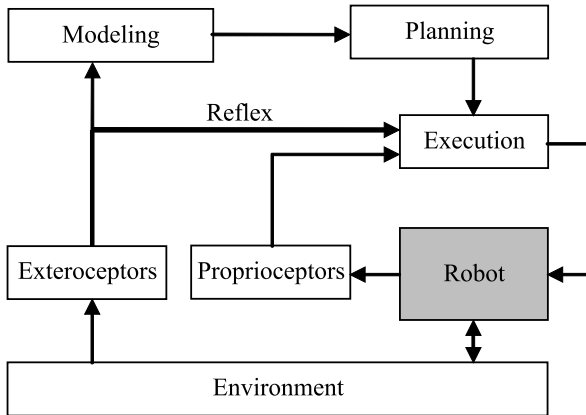
The conventional hierarchical control architecture that has been extensively used for mobile robots is shown in Fig. 4. The robot builds the model of itself using two types of sensors. The proprioceptive sensors, like shaft encoders, give information about the internal state of the



**Neuro-fuzzy Control of Autonomous Robotics, Figure 4**  
The low-level in the hierarchical architecture blindly executes the plan

robot. The exteroceptive sensors, like a sonar, provide information about the state of the environment. Using this information, a planning algorithm generates a plan that will perform the given task in the given environment. The instructions from the plan are blindly executed by the lower-level motor control layer. This layer uses proprioceptors for local feedback, but does not monitor the environment. If an obstacle is suddenly detected by the exteroceptors, it takes a long time for the robot to react. This happens because the stimulus has to pass through the higher layers first. Modeling and planning usually involve a lot of computation and take a long time. The response of the overall system to a new configuration of the environment is very slow. That is why robots implemented using this control hierarchy have to move with a very low speed in order to be able to avoid obstacles.

A solution to the above problem can be searched in the way humans react to some external stimuli, for example to pain. A simple spinal reflex can be traced in Fig. 3. When the pain receptor in the skin is excited, it fires a neuron in the LMN system, which in turn fires the LMN driving the flexor muscle. This operation removes the respective part of the body from danger, in a very fast and straightforward manner. There is no reaction expected from the higher layers of the nervous system before the motor neurons are fired. Still, the information from the skin receptor reaches the higher layers and a more intelligent measure is taken after a certain delay. The overall system now has a fast response to the environment. Reasoning about the necessary action can still be done and the movement can be corrected by the central nervous system through the direct inputs to the motor neurons. On mobile robots, the same type of reaction can be achieved by using low-level behaviors implemented in the execution layer. A simple but very important modification [25] has to be done to the structure in Fig. 4. The execution layer will not follow the planner commands blindly anymore, but will also receive



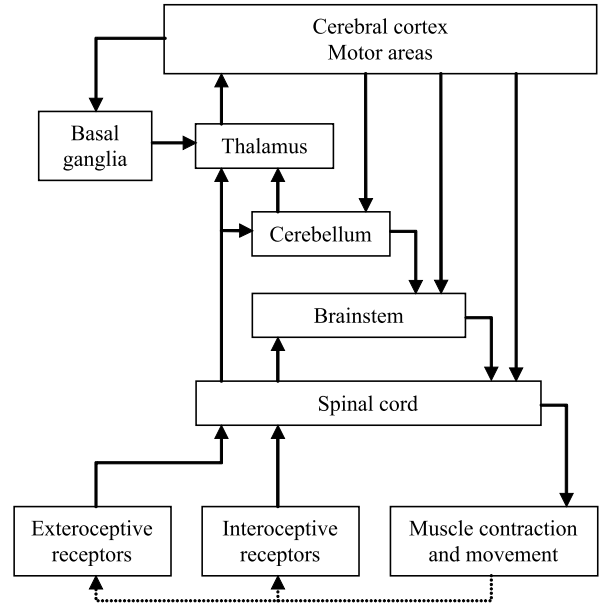
Neuro-fuzzy Control of Autonomous Robotics, Figure 5

The low-level execution layer can implement behaviors based on the sensed environment

information about the environment from the exteroceptive sensors, as shown in Fig. 5. The behaviors (that can be easily implemented using fuzzy logic) will have a fast reaction to the environment, because it is included now in the low-level feedback loop. The planner will only have to select the right behavior or to combine behaviors corresponding to the current goal and the current environment. The rapidity of decisions is not critical anymore for the higher levels.

From the overall organization of the human motor nervous system [10], shown in Fig. 6, and from the previous examples, a general principle can be extracted. First of all, (complete) sensorial data arrives to all the layers of control, starting from the lowest (the spinal cord) and up to the highest (the cerebral cortex). Each layer uses what information it needs and has certain autonomy in taking decisions, independent of the higher layer in the hierarchy. The latter will eventually correct or adjust the action initiated by the lower layers. The reaction speed is the fastest for the simple layers close to the effectors (of the order of hundreds of milliseconds) and decreases as one rises through the hierarchy and where structures become more complex.

In [3], Rodney A. Brooks has proposed a similar layered control architecture for mobile robots, organized into levels of competence. It is called the *subsumption architecture* and is based on incorporating lower levels of functionality into more general levels (Fig. 7). Each level of competence includes as a subset each earlier level of competence. Practically the higher layers impose additional restrictions on the behaviors implemented in the lower layers. The lowest-level layer is layer 0. It is implemented and tested by its own. Layer 1 is then built on top of layer 0. It can read information from layer 0 and can inject informa-



Neuro-fuzzy Control of Autonomous Robotics, Figure 6

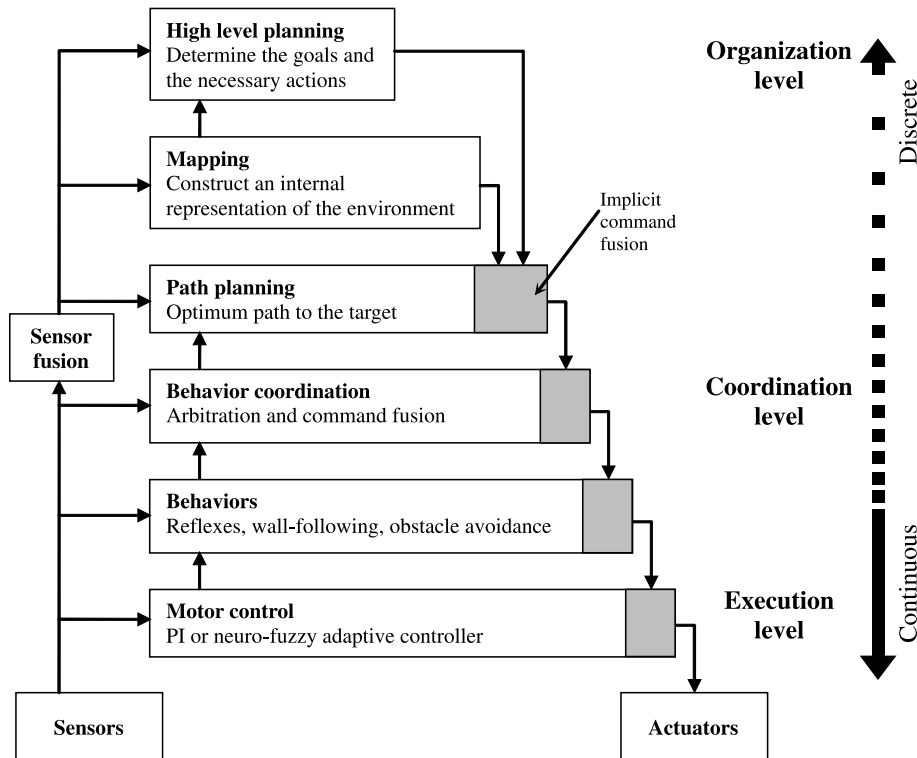
Summary of motor control in the human nervous system

tion into the internal interfaces of layer 0, suppressing the normal data flow. Level 1 of competence is achieved using layer 1 and the help of layer 0. The latter continues to run unaware of the fact that layer 1 interferes with its data paths when it wants to take control. Many layers of competence can be added one on top of the other using this mechanism. The functionality of the different layers for a mobile robot spawns across three main levels, as stated by Stephanou in [28] and by Lefebvre and Saridis in [16]: organization, coordination and execution.

The lower levels in the hierarchy have very fast reactions to sensor readings. They don't usually do complex processing on this information. The higher levels usually do sensor fusion and apply special algorithms to extract more abstract information about the environment. The decisions are taken with longer delays and with a much lower rate. It can be observed that there is a transition between almost continuous-time control at the lower levels, with high sample rates, and relatively slow discrete events at the highest levels. This transition can be exemplified with the difference between a fast PI controller for the wheels motors, and a discrete event controller for the implementation of the main plan of action.

### Map-Based Navigation

Any approach to control a dynamic system needs to use a model of the system to be controlled. In the case of a mobile robot, the system consists of the robot itself plus the



Neuro-fuzzy Control of Autonomous Robotics, Figure 7  
Subsumption architecture for robot control

environment in which it operates. It is easy to obtain the model of the robot on its own, but unfortunately the situation is different if we consider the robot to be embedded in a real-world, unstructured environment, that is characterized by uncertainty which is difficult to model or to quantify. The mapping problem is generally regarded as one of the most important problems in the pursuit of building truly autonomous mobile robots.

Map-based navigation calls upon three processes:

- Map learning – the process of memorizing the data acquired by the robot during exploration in a suitable representation.
- Localization – the process of deriving the current position of the robot within the map.
- Path planning – the process of choosing a sequence of actions in order to reach a goal, starting from the current position.

Map learning and localization are interdependent. Building a map requires the position of the robot to be estimated relative to the incomplete map while localization requires that the map exists. Path planning is rather independent of the other two and takes place once the map and the robot position are already available.

Various map representations have been used in the robotics literature. They are adequate or not depending on the task and the characteristics of the robot and the environment.

### Artificial Potential Fields

Artificial potential field methods for obstacle avoidance have gained increased popularity among researchers in the field of mobile robots. The idea of imaginary forces acting on a robot has been suggested by Andrews and Hogan [1] and Khatib [15]. The approach used to generate the artificial potential fields is to have obstacles exert repulsive forces onto the mobile robot, while the target applies an attractive force to it. The resultant force  $F$  determines the direction and speed of travel for the robot. The method is simple and elegant, yielding acceptable results with simple and quick implementations and without requiring many refinements. The moving robot can come in a variety of shapes and sizes. To simplify the development of the solution, the robot can be represented as a point while the obstacles and physical workspace boundaries are transformed by increasing their size with a value related to the dimension of the robot.

The potential is a scalar field whose negative gradient is a vector field of conservative forces. Any point of the force field will represent the resultant force obtained by summing the effects of the attraction force of the target and the repulsion forces of the obstacles:

$$\mathbf{F}(\mathbf{r}) = \mathbf{F}_{target}(\mathbf{r}) + \sum_{i=1}^n \mathbf{F}_{obs_i}(\mathbf{r}) = -\nabla V(\mathbf{r}) \quad (4)$$

For mobile robots applications, the potential fields are usually used in a 2-dimensional space:

$$\begin{aligned} F_x(x, y) &= -\frac{\partial V(x, y)}{\partial x} \\ F_y(x, y) &= -\frac{\partial V(x, y)}{\partial y} \end{aligned} \quad (5)$$

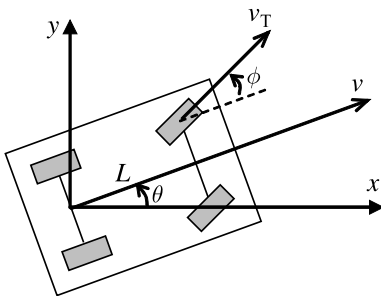
The trajectory is generated continuously for any position  $(x, y)$  of the robot following the direction of the steepest descent. It can be considered an optimization problem that searches for the point of minimum potential.

Let's consider the model of a car-like vehicle with no mass. It is a nonholonomic system, so it will most likely deviate from the ideal trajectory resulted from the potential field. However, the potential field implicitly provides a feedback mechanism. The  $\mathbf{F}$  force is generated for any position of the robot in such a way that it will return close to the ideal trajectory. The robot (Fig. 8) has the following dynamics:

$$\begin{aligned} \dot{x} &= v_T \cos \phi \cos \theta \\ \dot{y} &= v_T \cos \phi \sin \theta \\ \dot{\theta} &= \frac{v_T}{L} \sin \phi \end{aligned} \quad (6)$$

with  $(x, y)$  the position,  $\theta$  the heading angle,  $v_T$  the wheel speed,  $L$  the wheel base, and  $\phi$  the steering angle.

A potential field corresponding to the following environment is generated:



Neuro-fuzzy Control of Autonomous Robotics, Figure 8  
Model of a car-like robot

- Goal: a constant force will attract the robot to the target.

$$(x_G, y_G) = (10, 10)$$

$$r = \sqrt{(x_G - x)^2 + (y_G - y)^2}$$

$$F_{Gx}(x, y) = K_G \frac{x_G - x}{r}, \quad F_{Gy}(x, y) = K_G \frac{y_G - y}{r}$$

- Circular obstacles: a force similar to the electrostatic force will repel the robot from the obstacles.

$$(x_1, y_1) = (3, 3) \text{ for the first obstacle}$$

$$(x_2, y_2) = (7, 2) \text{ for the second obstacle}$$

$$F_{ix}(x, y) = -K_i \frac{x_i - x}{(\max\{(r - a_i), b_i\})^2},$$

$$F_{iy}(x, y) = -K_i \frac{y_i - y}{(\max\{(r - a_i), b_i\})^2},$$

where  $r$  is the distance between the robot and the goal or the obstacle,  $a_i$  is the radius of the obstacle and  $b_i$  limits the relative height of the obstacle. The total force that acts on the robot is

$$F_x = F_{Gx} + F_{1x} + F_{2x}$$

$$F_y = F_{Gy} + F_{1y} + F_{2y}$$

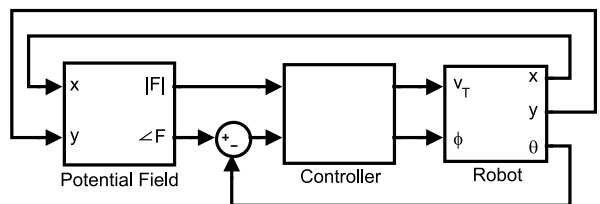
and the angle of the force as seen from the robot is

$$\alpha = \tan^{-1} \left( \frac{F_y}{F_x} \right)$$

with corrections for the correct quadrant.

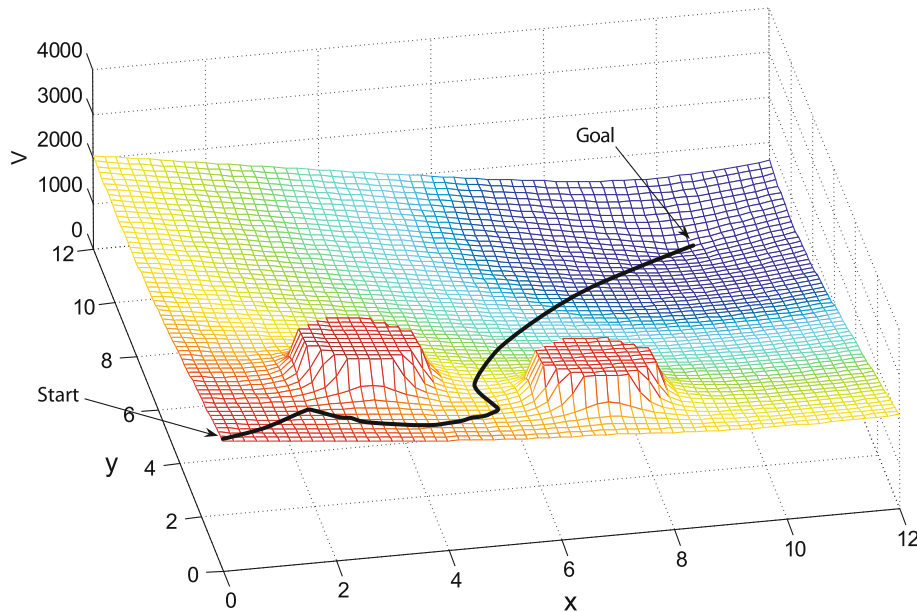
The initial position for the robot is  $(x_0, y_0) = (0, 0)$  and the initial heading angle is  $\phi_0 = \pi/6$ . The speed  $v_T$  is often kept constant. The steering angle may be generated by the controller based on the difference between the heading angle  $\theta$  and the angle  $\alpha$  of the resultant force acting on the robot (Fig. 9). In this example, a proportional controller is implemented:

$$\phi = K(\theta - \alpha)$$



Neuro-fuzzy Control of Autonomous Robotics, Figure 9  
Closed-loop control based on the potential field





Neuro-fuzzy Control of Autonomous Robotics, Figure 10  
The trajectory of the robot using the potential field

### Problems Associated with Potential Field Methods

It can be seen from the potential field representation (Fig. 10) that the robot will follow the direction of the steepest descent, which is the direction of the resultant force. There are a few issues that can prevent the robot to reach the target:

- Trap situations due to local minima
- No passage between closely spaced obstacles
- Oscillations in the presence of obstacles or narrow passages

Some of the problems can be solved by modifying the height or dimension of the obstacles, by changing the type of force used to represent them or by using an intelligent controller in the structure from Fig. 9.

### Fuzzy Logic and Potential Fields

A good example of combining fuzzy logic and potential fields for path planning can be found in the work of Valavanis et al. [30] that we will present in the following pages. They have implemented a two-layered fuzzy logic inference engine and a discrete type of potential field for real-time mobile robot navigation in a 2-D dynamic environment. The first layer of the fuzzy logic inference engine performs sensor fusion from sensor readings into a fuzzy variable, *collision*, providing information about possible

collisions in four directions: *front*, *back*, *left* and *right*. The second layer guarantees collision avoidance with dynamic obstacles while following the trajectory generated by the Electrostatic Potential Field (EPF). The main idea is to combine planned and reactive behavior. Given a 2-D environment (with initial information from potentially existing environment a priori maps and on-line sonar sensor data), the EPF plans the initial trajectory and starts executing it. Once the object detection module (working in parallel with the EPF) detects using sensor readings a “high collision possibility”, it forces the motion control module to “forget” the initial EPF path, take corrective actions in terms of robot steering and robot speed to avoid the collision, until new sensor readings dictate a “low” or “not-possible” collision possibility. Then, the motion control module takes into account the initial trajectory as computed at this time instant by the EPF planner. The EPF planner is invoked every time the environment map is updated.

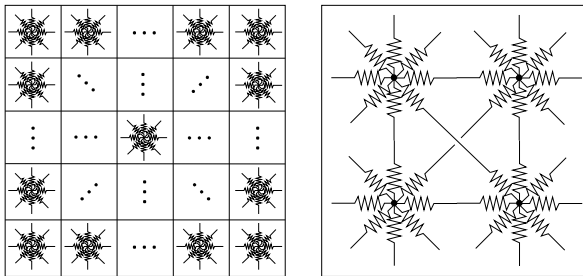
The solution to the navigation problem may be compared to the flow of electric current within a sheet of conducting material. Obstacles are mapped into a discrete resistor network. The path of minimum resistance within the circuit corresponds with a path of minimum occupancy within the environment. The algorithm used to create the natural potential field follows three steps:

- Obtain an occupancy map of the environment
- Create the resistor network

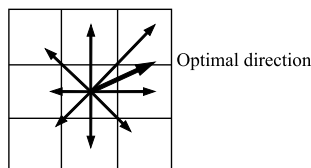
- Calculate the values of the resistors in the network to obtain the potential field

Each cell in the occupancy map is replaced with a set of eight resistors (Fig. 11), each connected at a central point. The only exceptions are the cells on the outside edges and corners with five and respectively three resistors connected. The resistors values depend on the corresponding cell in the occupancy map. A minimum resistance path in the network determines a maximum potential drop. Reversing the mapping generates an optimal path in the environment corresponding to a minimum occupancy path. The path always begins at the highest potential (initial vehicle position), and ends at the lowest potential (final destination).

Consider an environment map which contains obstacles of various shapes and sizes. The initial position of the robot is  $q_0$  and the destination point is  $q_f$ . Assume a square, bounded region centered about  $q_0$  which includes  $q_f$  and can be divided into an  $n \times n$  grid,  $X$ . The grid is discretely represented by the occupancy matrix  $O$ , where the value of each entry is the percentage of the area of the grid cell occupied by obstacles of the environment map. To determine a desired direction of travel from the EPF, a vector is associated with each cell connected to the cell containing  $q_0$  with magnitude equal to the amount of current flowing through the specified branch. Figure 12 shows the vectors created through a solution of the resistor network.



Neuro-fuzzy Control of Autonomous Robotics, Figure 11  
An  $n$  by  $n$  resistor network and node detail



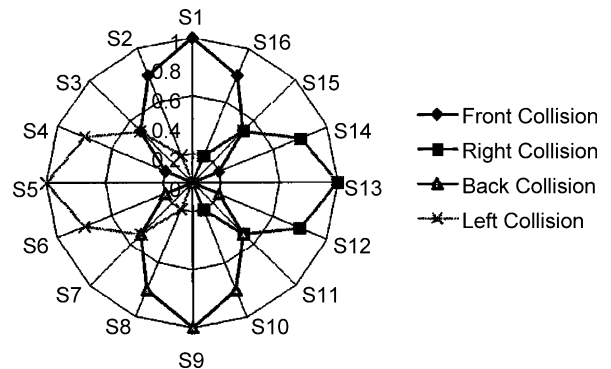
Neuro-fuzzy Control of Autonomous Robotics, Figure 12  
Vector representation of the EPF solution

The magnitudes of the current along each branch are shown at the end of the associated vector. The geometric sum of these forces results in a single vector pointing in the direction of the optimal path, as shown in Fig. 12.

The control of the robot is obtained using a two-layered FL interference engine. The first layer handles obstacle detection and performs sensor data fusion from sonar readings. There are two input fuzzy variables: *sensor\_direction* and *sensor\_distance*. The variable *sensor\_direction* has four different values that describe the sensor's membership in four cardinal relative directions: *front\_collision*, *left\_collision*, *back\_collision*, and *right\_collision*. Each of the collision values is represented as a fuzzy variable with values *not\_possible*, *possible*, and *high*. The second layer has three inputs. The first receives the output of the first inference, representing the immediate collision possibility. The second input receives the output of the artificial potential field and the third receives the speed of the robot. There are two outputs that are used to control the movement of the robot: change of speed and steering.

A mobile robot, the Nomad 200, with a ring of 16 sonar sensors that cover a full 360 circle is used for experimental purposes. The sonar sensors are fixed to the body of the robot. Therefore, *sensor\_direction* is a function only of the number of the sonar. Sonar S1 always belongs to *front\_collision* with weight 1.0 and sonar S2 always belongs to *front\_collision* with weight 0.8 and *left\_collision* 0.2, as shown in Fig. 13.

To simplify the implementation, *sensor\_direction* has not been used but instead each sonar has been assigned to at least one but no more than two values of the variable collision. The rule base of the obstacle detection controller contains rules of the following two types:



Neuro-fuzzy Control of Autonomous Robotics, Figure 13  
Sonar arrangement and relative importance to collision detection [30]

- R1) IF  $d_i$  is  $LD^{(k)}$  THEN  $c_j$  is  $LC^{(k)}$ ;  
 R2) IF  $d_i$  is  $LD^{(k)}$  AND  $d_{i-1}$  is  
 $LD^{(k)}$  AND  $d_{i+1}$  is  $LD^{(k)}$  THEN  $c_j$  is  $LC^{(k)}$

where

- $k$  is the rule number,
- $d_i$  represents the readings of sensor  $i$ ,
- $LD^{(k)}$  is a linguistic value of the term set  $D = \{Close, Near, Far\}$ ,
- $c_j$  is the collision of type  $j$  ( $j \in \{Front, Left, Right, Back\}$ ) and
- $LC^{(k)}$  is a linguistic value of the term set  $C = \{Not-possible, Possible, High\}$ .

The whole rule-base is presented in [13]. Some of the rules for the collision of type  $j=front$  are

- IF  $d_1$  is *Close* THEN *front* collision is *High*.  
 IF  $d_1$  is *Far* AND  $d_2$  is *Far* AND  $d_{16}$  is *Far* THEN *front*  
 collision is *Not-Possible*.  
 IF  $d_2$  is *Near* THEN *front* collision is *Possible*.  
 IF  $d_1$  is *Near* AND  $d_2$  is *High* AND  $d_{16}$  is *Far* THEN  
*front* collision is *Possible*.  
 IF  $d_1$  is *Near* AND  $d_2$  is *Far* AND  $d_{16}$  is *Far* THEN  
*front* collision is *Possible*.

The mathematical meaning of the  $k$ th single antecedent rule (type R1) is given as a fuzzy relation  $R^{(k)}$  on  $D \times C$ , which in the membership functions domain is

$$\mu_{R^{(k)}}(d_i, c_j) = \min[\mu_{LD^{(k)}}(d_i), \mu_{LC^{(k)}}(c_j)] \quad (7)$$

The membership function of the obstacle's position and distance from the robot is computed by the max-min composition between the sensor readings, which represent the distance from the obstacles and the fuzzy relation described by (7). The second layer fuzzy controller takes as input the variables: *collision*, *angle\_error*, and *speed* and generates the control *change of speed* and *steering*. The potential field generates a heading directive based upon the sum of forces approach described in the previous section. The heading relative to the current heading of the robot is given as the *angle\_error*  $\theta$ . This angle represents the angle in which the vehicle should point in order to follow the desired path to the goal point. Since the desired heading given by the potential field already considers some reactive navigation, the second-layer fuzzy inference should implement the steering necessary to reduce the *angle\_error* to zero, only changing the desired heading if collision in the vicinity of the desired path is possible.

Reaction-directed behaviors are controlled through analysis of the collision of the current situation. For example, IF *collision* is *front possible* AND *speed* is *normal*

THEN *change of speed* is *decelerate slow*. A similar rule, which examines the *angle\_error* as well as all values of *collision*, generates the control *steering*. The rules of the second fuzzy module can be described compactly as follows:

- IF  $c_j$  is  $LC^{(k)}$  AND  $\theta$  is  $L\Theta^{(k)}$  AND  $v$  is  $LV^{(k)}$  THEN  $s$   
 is  $LS^{(k)}$  AND  $dv$  is  $LDV^{(k)}$

where

- $k$  is the rule number,
- $c_j$  is collision of type  $j$ ,
- $\theta$  is the steering angle error,
- $v$  is the speed of the robot,
- $s$  is the steering angle correction,
- $dv$  is the change of speed, and
- $LC^{(k)}$ ,  $L\Theta^{(k)}$ , and  $LV^{(k)}$ ,  $LS^{(k)}$ ,  $LDV^{(k)}$  are the linguistic values of  $c_j$ ,  $\theta$ ,  $v$ ,  $s$  and  $dv$ .

In all the rules, AND is the min operator. The steering angle error  $\theta$  is computed by continuously comparing the desired steering angle,  $\theta_{desired}$ , i. e., the angle that guides the robot to the target point given, with the actual steering angle,  $\theta_{actual}$ . The steering error is important when the robot detects no collision on the target path. In case of possible collision, the information about the error  $\theta$  becomes of less importance and the rules that contain such information are firing with smaller strength than the rules which perform collision avoidance. The generic mathematical expression of the  $k$ th navigation rule is

$$\mu_{R^{(k)}}(c_j, \theta, v, s, dv) = \min[\mu_{LC^{(k)}}(c_j), \mu_{L\Theta^{(k)}}(\theta), \mu_{LV^{(k)}}(v), \mu_{LS^{(k)}}(s), \mu_{LDV^{(k)}}(dv)] \quad (8)$$

The overall navigation output is given by the max-min composition and in particular is

$$\mu_N^*(s, dv) = \max_{c_j, \theta, v} \min[\mu_{AND}^*(c_j, \theta, v), \mu_R(c_j, \theta, v, s, dv)] \quad (9)$$

where

$$\mu_R(c_j, \theta, v, s, dv) = \bigcup_{k=1}^K \mu_{R^{(k)}}(c_j, \theta, v, s, dv)$$

and  $\mu_{AND}^*(c_j, \theta, v)$  is the minimum of the fuzzified sonar readings. The navigation action dictates change in robot speed and/or steering correction as it comes out from a defuzzification formula, which calculates the center of the area covered by the membership function computed from (9). Part of the rule base is shown in Table 1.

Neuro-fuzzy Control of Autonomous Robotics, Table 1

Rule base for the fuzzy inference engine implemented in [30]

Input variables						Output variables	
Collision				Speed	Steering angle error	Change of speed	Steering angle correction
Front	Left	Back	Right				
N-P	N-P	N-P	N-P	Low	RB	NC	TLF
N-P	N-P	N-P	N-P	Normal	R	NC	TL
N-P	N-P	N-P	N-P	Normal	Zero	ACC-F	NC
N-P	High	N-P	N-P	High	LS	DEC-SL	TRS
N-P	N-P	N-P	High	Low	LB	NC	TRS
High	High	N-P	N-P	Low	Zero	NC	TRF
High	N-P	High	High	Normal	RS	DEC-F	TLF
N-P	High	High	N-P	Normal	R	DEC-SL	NC
N-P	High	N-P	High	High	RB	DEC-SL	NC
High	High	High	N-P	High	Zero	DEC-F	TRF
High	N-P	High	N-P	Low	LS	NC	TR

### Neuro-Fuzzy Assisted EKF for SLAM Problems

The most successful robot navigation algorithms have been derived from a probabilistic perspective [29], which takes into account vehicle motion, terrain uncertainty and sensor noise. The interest in the estimation of an autonomous robot's location and that of its surroundings, known as Simultaneous Location and Map Building (SLAM), is evident. The goal of an autonomous vehicle performing SLAM is to build a map consisting of environment features (landmarks) incrementally, by using the uncertain information extracted from its sensors, while simultaneously using that map to localize itself with respect to the reference coordinate frame. Extended Kalman Filter (EKF) [17] has been a popular choice for SLAM [8]. This approach estimates and stores the robot position and orientation and the feature positions within the map of the environment as a complete state-vector. The uncertainties in these estimates are stored as error covariance matrices. It has been shown previously that the performance of an EKF process depends largely on the accuracy of the knowledge of the process noise covariance matrix ( $\mathbf{Q}$ ) and measurement noise covariance matrix ( $\mathbf{R}$ ). An incorrect a priori knowledge of  $\mathbf{Q}$  and  $\mathbf{R}$  can lead to performance degradation or even divergence [11,20].

In real-world applications, the  $\mathbf{Q}$  and  $\mathbf{R}$  matrices may not be accurately known. In [5] Chatterjee and Matsuno suggest the use of fuzzy logic techniques for the adaptation of the statistical assumption of the EKF caused by unknown or possible changed sensor noise characteristics. Only the adaptive estimation of  $\mathbf{R}$  is considered, which is done at each iteration. We will present the main aspects of their implementation as an example on how fuzzy logic

can improve some important methods used for robot navigation, even if their intrinsic nature is completely different.

The algorithm assumes that the features do not move and are modeled as static 2D points, and that Gaussian noise is affecting the system states and the measurements (represented by the range  $r$  and the bearing  $\theta$ ).

The EKF is used when the process state transition is modeled by a nonlinear function or when there is a nonlinear relation between the measurements and the states of the system:

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) + \mathbf{q}_k \quad (10)$$

$$\mathbf{z}_{k+1} = \mathbf{h}(\mathbf{x}_{k+1}) + \mathbf{r}_{k+1} \quad (11)$$

where  $\mathbf{x}_k \in \mathbf{R}^n$  is the state vector of the process with the corresponding covariance matrix  $\mathbf{P}_k$ ,  $\mathbf{z}_k \in \mathbf{R}^m$  is the measurement vector and  $\mathbf{u}_k$  is the control input. The random variables  $\mathbf{q}_k$  and  $\mathbf{r}_k$  represent the process and measurement noise (respectively). They are assumed to be independent (of each other), white, and with normal probability distributions

$$p(\mathbf{q}) \sim N(0, \mathbf{Q})$$

$$p(\mathbf{r}) \sim N(0, \mathbf{R})$$

For the SLAM algorithm, the coordinates of the features are considered states of an augmented system. The total state vector  $\mathbf{x}$  will contain the states of the vehicle  $\mathbf{x}_v$  (position and orientation in the reference frame) and the states of the already observed features  $\mathbf{x}_m$  (a vector of variable length). The mean estimate of the total state vector and the corresponding total error covariance matrix  $\mathbf{P}$  are the

following:

$$\hat{\mathbf{x}} = [\hat{\mathbf{x}}_v^T \quad \hat{\mathbf{x}}_m^T]^T \quad (12)$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_v & \mathbf{P}_{vm} \\ \mathbf{P}_{vm}^T & \mathbf{P}_m \end{bmatrix} \quad (13)$$

$$\hat{\mathbf{x}}_v = [\hat{x}_v \quad \hat{y}_v \quad \hat{\phi}_v]^T \quad (14)$$

$$\hat{\mathbf{x}}_m = [\hat{x}_1 \quad \hat{y}_1 \quad \dots \quad \hat{x}_n \quad \hat{y}_n]^T \quad (15)$$

The state transition in (10) is modeled by the following nonlinear function that keeps the feature states constant:

$$\mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) = \begin{bmatrix} \mathbf{f}_v(\mathbf{x}_{v_k}, \mathbf{u}_k) \\ \mathbf{x}_{m_k} \end{bmatrix} \quad (16)$$

The map is defined from the position estimates of the static features as they appear in (15). The number of features  $n$  is variable. Initially the algorithm is started with  $n = 0$ . They are added in subsequent iterations, when the robot starts moving and observations are carried out. The control inputs to the system are also considered to be random variables, because they are measurements (e. g. from odometers) affected by errors. They are modeled as Gaussian variations from their nominal values with normal probability distribution

$$p(\mathbf{u}) \sim N(0, \mathbf{U})$$

The following equations implement the discrete EKF. A complete description can be found in [31].

*EKF Time Update (“Predict”) equations:*

$$\hat{\mathbf{x}}_{k+1}^- = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k) \quad (17)$$

$$\mathbf{P}_{k+1}^- = \nabla \mathbf{f}_{\mathbf{x}_k} \mathbf{P}_k \nabla \mathbf{f}_{\mathbf{x}_k}^T + \nabla \mathbf{f}_{\mathbf{u}_k} \mathbf{U}_k \nabla \mathbf{f}_{\mathbf{u}_k}^T \quad (18)$$

$$\nabla \mathbf{f}_{\mathbf{x}_k} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}_k} \right|_{(\hat{\mathbf{x}}_k, \hat{\mathbf{u}}_k)}$$

$$\nabla \mathbf{f}_{\mathbf{u}_k} = \left. \frac{\partial \mathbf{f}}{\partial \mathbf{u}_k} \right|_{(\hat{\mathbf{x}}_k, \hat{\mathbf{u}}_k)}$$

*EKF Kalman Gain equations:*

$$\mathbf{v}_{i,k+1} = \mathbf{z}_{k+1} - \mathbf{h}_i(\hat{\mathbf{x}}_{k+1}^-) \quad (19)$$

$$\mathbf{S}_{i,k+1} = \nabla \mathbf{h}_{\mathbf{x}_{k+1}} \mathbf{P}_{k+1}^- \nabla \mathbf{h}_{\mathbf{x}_{k+1}}^T + \mathbf{R}_k \quad (20)$$

$$\mathbf{K}_{i,k+1} = \mathbf{P}_{k+1}^- \nabla \mathbf{h}_{\mathbf{x}_{k+1}}^T \mathbf{S}_{i,k+1}^{-1} \quad (21)$$

$$\nabla \mathbf{h}_{\mathbf{x}_{k+1}} = \left. \frac{\partial \mathbf{h}_i}{\partial \mathbf{x}_k} \right|_{\hat{\mathbf{x}}_{k+1}^-}$$

*EKF Measurement Update (“Correct”) equations:*

$$\hat{\mathbf{x}}_{k+1}^+ = \hat{\mathbf{x}}_{k+1}^- + \mathbf{K}_{i,k+1} \mathbf{v}_{i,k+1} \quad (22)$$

$$\mathbf{P}_{k+1}^+ = \mathbf{P}_{k+1}^- - \mathbf{K}_{i,k+1} \mathbf{S}_{i,k+1} \mathbf{K}_{i,k+1}^T \quad (23)$$

The features are measured relative to the observer in terms of their range  $r$  and bearing  $\theta$ . The measurement vector is given as

$$\mathbf{z} = [r \quad \theta]^T \quad (24)$$

and the measurement noise is described by the covariance matrix

$$\mathbf{R} = \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\theta^2 \end{bmatrix} \quad (25)$$

where it was assumed that there is no cross-correlation between the range and the bearing measurements.

In the previous equations,  $\mathbf{v}_i$  is the innovation of the observation corresponding to the  $i$ th landmark,  $\mathbf{S}_i$  is the covariance matrix of the innovation and  $\mathbf{h}_i$  is a function that maps the  $(\hat{x}_i, \hat{y}_i)$  estimated coordinates of the  $i$ th landmark to range and bearing coordinates relative to the current robot position and orientation.

The approach for adapting the EKF used in [5] by Chatterjee and Matsuno is based on the innovation adaptive estimation. The method was originally proposed in [20] and combined with fuzzy logic in [19]. The basic idea behind the method is to determine innovation of the observation at every  $k$ th instant (the discrepancy between a new measurement  $\mathbf{z}_k$  and its corresponding estimation  $\hat{\mathbf{z}}_k$ ) and to use the new information to correct the estimate already made. The adaptation strategy has the objective of reducing the mismatch between the theoretical covariance of the innovation sequences ( $\mathbf{S}_k$ ) and the corresponding actual covariance of the innovation sequences ( $\hat{\mathbf{C}}_{Innk}$ ).  $\mathbf{S}_k$  is calculated using (20) and  $\hat{\mathbf{C}}_{Innk}$  is calculated as

$$\hat{\mathbf{C}}_{Innk} = \mathbf{v}_k \mathbf{v}_k^T \quad (26)$$

When an observation step is carried out, there will be multiple landmarks visible simultaneously. The algorithm considers batch measurements of the form

$$\mathbf{z} = [r_1, \quad \theta_1, \quad \dots \quad r_n, \quad \theta_n]^T \quad (27)$$

and as a result the innovation vector is

$$\mathbf{v} = [\mathbf{v}_1^T \quad \mathbf{v}_2^T \quad \dots \quad \mathbf{v}_n^T]^T \quad (28)$$

The matrices involved in the algorithm are also constructed for batch mode measurements and have corresponding dimensions. The EKF algorithm performs better



update steps for SLAM if the innovation vector  $v$  contains more observations simultaneously. The mismatch of the innovation sequence covariance at the  $k$ th instant is

$$\Delta \hat{C}_{Innk} = \hat{C}_{Innk} - S_k$$

The objective is to minimize this mismatch by employing a one input – one output neuro fuzzy system for each diagonal element of the  $\Delta \hat{C}_{Innk}$  matrix. The fuzzy rules will adapt the sensor statistics defined by the  $R$  matrix so that there will be a reduction in the  $\Delta \hat{C}_{Innk}$  mismatch. The adaptation algorithm is run during the EKF Measurement Update step. It does the following actions:

- Determine the set of visible landmarks from the current robot position and do the range-bearing measurements for each of them to obtain  $z$  as in (27).
- Predict the range-bearing observations on the basis of the augmented total step vector and compute the augmented innovation sequence using (19) and (28) for batch processing.
- Compute the augmented noise covariance matrix for the measurements using the original  $(2 \times 2)$   $R$  matrix, the augmented observation model  $h$  and the augmented  $S$  for batch-mode situations.
- Update the a-posteriori state estimate vector and the error covariance matrix using (22) and (23).
- Compute  $\hat{C}_{Innk}$ ,  $\Delta \hat{C}_{Innk}$  and normalize the diagonal entries of  $\Delta \hat{C}_{Innk}$  corresponding to the range mismatch or to the bearing mismatch with respect to the highest value of the respective measurement on the diagonal.
- Apply each  $\Delta \hat{C}_{Innk}(j, j)$  element from the diagonal as an input to the *neuro-fuzzy system* and obtain the corresponding  $\Delta R(j, j)$  output.
- Determine  $\Delta \sigma_r^2$  and  $\Delta \sigma_\theta^2$  as means of those  $\Delta R(j, j)$  entries which correspond to range measurements and bearing measurements, respectively.

- Adapt the original  $(2 \times 2)$   $R$  matrix in the following manner:

$$R_k = R_{k-1} + \begin{bmatrix} \Delta \sigma_r^2 & 0 \\ 0 & \Delta \sigma_\theta^2 \end{bmatrix}. \quad (29)$$

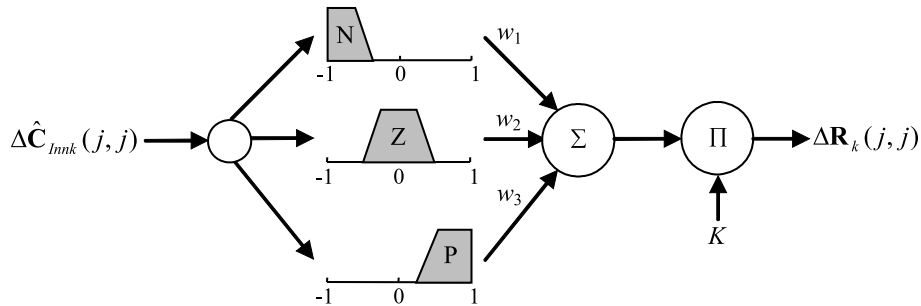
From the algorithm, it can be seen that the neuro-fuzzy system (NFS) employs a nonlinear mapping between the input  $\Delta \hat{C}_{Innk}(j, j)$  and the output  $\Delta R(j, j)$ . For each diagonal element of innovation covariance mismatch there will be a corresponding adaptation value recommended for the diagonal element of the augmented noise covariance matrix  $R$ . Because the same NFS is applied for every diagonal element, the input to the NFS is normalized. The NFS is implemented by three IF-THEN rules:

$$\begin{aligned} \text{IF } \Delta \hat{C}_{Innk}(j, j) \text{ is } N \text{ THEN } \Delta R(j, j) &= w_1 \\ \text{IF } \Delta \hat{C}_{Innk}(j, j) \text{ is } Z \text{ THEN } \Delta R(j, j) &= w_2 \\ \text{IF } \Delta \hat{C}_{Innk}(j, j) \text{ is } P \text{ THEN } \Delta R(j, j) &= w_3 \end{aligned}$$

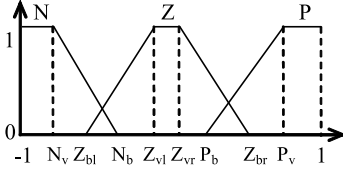
where  $w_1$ ,  $w_2$  and  $w_3$  indicate the amount of fuzzy adaptation recommended in the form of a diagonal element of the  $\Delta R$  matrix that depends on the nature of the mismatch in the corresponding diagonal element of the  $\Delta \hat{C}_{Innk}$  matrix. The neuro-fuzzy adaptation system is represented in Fig. 14 and the fuzzy membership functions in Fig. 15.

The  $\Delta \hat{C}_{Innk}(j, j)$  input variable is fuzzified using three Membership Functions (MFs)  $\mu_i$ : negative (N), zero (Z) and positive (P). The output of the defuzzification layer is calculated as a weighted average of all its inputs. Finally, the output layer performs a suitably scaling for the defuzzified output. The complete input-output relationship is the following:

$$\Delta R_k(j, j) = K \cdot \frac{\sum_{i=1}^3 \mu_i(\Delta \hat{C}_{Innk}(j, j)) \cdot w_i}{\sum_{i=1}^3 \mu_i(\Delta \hat{C}_{Innk}(j, j))} \quad (30)$$



Neuro-fuzzy Control of Autonomous Robotics, Figure 14  
Architecture of the neuro-fuzzy system



**Neuro-fuzzy Control of Autonomous Robotics, Figure 15**  
Parametrization of the membership functions

The neuro-fuzzy model has to be trained in order to determine the suitable parameters of the membership functions, the output consequence singletons and the output gain. Unfortunately, the training can not be completed in the conventional supervised mode because the desired output for a given input is not known. A stochastic global optimization algorithm is needed for training in an unsupervised manner. The authors have chosen to use Particle Swarm Optimization (PSO), which is a relatively new algorithm based on the swarm behavior of birds and fishes. A detailed description of the algorithm can be found in [6]. The training is done as a high-dimension metaheuristic problem with the objective of minimizing a fitness function  $f_{\text{fit}}(x_1, x_2, \dots, x_n)$  on the basis of the variables  $x_1, x_2, \dots, x_n$ .

The “particles” in a PSO problem are several candidate solutions of  $x_1, x_2, \dots, x_n$ . At each iteration, the suitability of each solution is evaluated. For the training of the neuro-fuzzy system, the particles are formed as 12 component vectors

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_{12} \end{bmatrix} = \begin{bmatrix} N_v & N_b & Z_{bl} & Z_{vl} & Z_{vr} & Z_{br} & P_b & P_v & w_1 & w_2 & w_3 & K \end{bmatrix}$$

Each particle is characterized by vectors denoting its position  $\mathbf{x}_i$  and its velocity  $\mathbf{v}_i$ , which are adjusted at each time step in order to minimize the fitness function  $f_{\text{fit}}(x_1, x_2, \dots, x_n)$ . The traditional PSO algorithm has a quick convergence during the early phase of the training procedure, but has problems in finding the precise solution to the problem. To improve the behavior of the PSO algorithm, the variation for the speed is kept high in the initial passes and is gradually decreased as iterations progress.

The formulation of the fitness function has to be in concordance with the objective of the neuro-fuzzy assistance to the EKF. The discrepancy between the theoretical covariance and the actual covariance of the innovation sequence over the entire range of observation instants has to be minimized. For each iteration there are  $N_{\text{obs}}$  observations considered. The fitness function calculates the mean-square value of all the diagonal entries of  $\Delta \hat{\mathbf{C}}_{\text{Innk}}$  for each

observation instant and computes a mean of these values for the entire batch of observations:

$$f_{\text{fit}} = \frac{1}{N_{\text{obs}}} \sum_{n_{\text{obs}}=1}^{N_{\text{obs}}} \left( \frac{1}{J_{Cnobs}} \sum_{j=1}^{J_{Cnobs}} [\Delta \mathbf{C}_{\text{Innk}}(j, j)]^2 \right) \quad (31)$$

where  $J_{Cnobs}$  is the number of diagonal elements of the  $\Delta \hat{\mathbf{C}}_{\text{Innk}}$  matrix ( $2 \times$  the number of observed features) when the observation  $n_{\text{obs}}$  takes place.

The positions of each of the particles at the end of each iteration have to comply with several constraints due to the specific shapes that have to be obtained for the fuzzy membership functions (in this case trapezoidal) and also to ensure some overlapping between consecutive MFs. All the points from the universe of discourse have to be covered by at least one MF. The constraints are implemented using the following type of rules:

IF ( $N_b < N_v$ ) THEN  $N_b = N_v$ ,  
IF ( $Z_{vl} < Z_{bl}$ ) THEN  $Z_{vl} = Z_{bl}, \dots$  etc.

The performance of the algorithm was compared with a conventional EKF SLAM where the noise covariance matrices  $\mathbf{Q}$  and  $\mathbf{R}$  are kept constant. Both algorithms start with the same values for the matrices, but the neuro-fuzzy algorithm adapts  $\mathbf{R}$  continuously. The EKF-based SLAM algorithm performs very well when the sensor statistics are known. The performance degrades significantly and becomes highly unreliable when the statistics are wrongly considered, especially for a high number of landmarks. The neuro-fuzzy assistance improves the situation dramatically, helping the EKF to maintain a stable performance over several environment situations while having wrong knowledge of the sensor statistics.

The training of the neuro-fuzzy system was done in offline mode using the data gathered by the robot for a given configuration of the environment. Once the NFS parameters were determined, experiments were done for robot navigation through waypoints for several configurations of landmarks.

## Behaviors

The operational characteristics of unmanned vehicles include the following:

- Perception: acquire knowledge about the environment using sensors and extracting meaningful information to be used in later tasks
- Intelligence: operate for a considerable amount of time without human intervention while accomplishing useful tasks
- Action: in general, move between certain points.

One of the most common applications of fuzzy logic for autonomous robotics is to implement individual behaviors. Considering the environment uncertainty that is difficult if not impossible to model, effective control algorithms for autonomous navigation should imitate the way humans are operating various vehicles. Fuzzy logic allows a suitable knowledge representation of inherently vague notions achieved through IF-THEN rules. These rules contain linguistic information that describes the problem in a simple and fast manner. In many applications of fuzzy logic, a mathematical model of the dynamics of the vehicle is not needed. The only requirement for the design of the inference engine is the heuristic control knowledge related to the specific problem, usually obtained from a human who knows how to handle the system. Tuning is done by trial and error methods. Because of their interpolative nature, fuzzy controllers generate smooth movement for the robots and provide robustness and a graceful degradation of performance when confronted with noise and other type of errors in the data obtained from the sensors.

### Feedback Linearization Using Neural Networks

The lowest level of control on mobile robots is the execution level (Fig. 7). The higher levels generate a desired trajectory for the robot. The motor control unit has to do force control, position control or speed control. In some cases, a simple PID controller can be used with good results. But usually the exact model of the system is unknown and even worse, it is also nonlinear. A simple linear controller may not be robust or might not have a good performance. Neural networks can be used to build “model-free” controllers that offer a powerful and robust alternative to adaptive control. Mimicking the functions of human processes, these controllers learn about the systems they are controlling on-line, and thus automatically improve their performance.

Feedback linearization techniques offer a widely applicable set of design tools that are useful for broad classes of nonlinear systems. They function by basically converting the nonlinear problem into a related linear controls design problem. The performance is far exceeding that of classical linearization controllers based on Jacobian linearization techniques because no approximation is involved in feedback linearization design.

Lewis et al. [18] have done feedback linearization for a robotic manipulator using a neural network to estimate the nonlinearities of the system. The method can be applied to a large class of nonlinear systems.

The robot manipulator has dynamics

$$M(q)\ddot{q} + V_m(q, \dot{q})\dot{q} + F(\dot{q}) + G(q) + \tau_d = \tau \quad (32)$$

To make the robot manipulator follow a prescribed desired trajectory  $q_d(t)$ , define the tracking error  $e(t)$  and filtered tracking error  $r(t)$  by

$$e(t) = q_d - q \quad (33)$$

$$r = \dot{e} + \Lambda e \quad (34)$$

with  $\Lambda > 0$  a positive definite design parameter matrix. The robot dynamics are expressed in terms of the filtered error as

$$M\dot{r} = -V_m r + f(x) + \tau_d - \tau \quad (35)$$

where the nonlinear robot function is defined as

$$f(x) = M(q)(\ddot{q}_d + \Lambda\dot{e}) + V_m(q, \dot{q})(\dot{q}_d + \Lambda e) + F(\dot{q}) + G(q) \quad (36)$$

One may define

$$x \equiv [e^T \quad \dot{e}^T \quad q_d^T \quad \dot{q}_d^T \quad \ddot{q}_d^T]^T.$$

According to the universal approximation property of the neural network, there is a two-layer NN such that

$$f(x) = W^T \sigma(V^T x) + \varepsilon \quad (37)$$

with the approximation error  $\varepsilon$  bounded on a compact set by a known bound.  $W$  and  $V$  are unknown ideal target weights that give good approximation to  $f(x)$ . It is only necessary to know that they exist.

The control input is selected as

$$\tau = \hat{W}^T \sigma(\hat{V}^T x) + K_v r - \dot{v} \quad (38)$$

with  $v(t)$  a function that provides robustness:

$$v(t) = -K_z (\|\dot{Z}\|_F + Z_B) r, \quad (39)$$

where

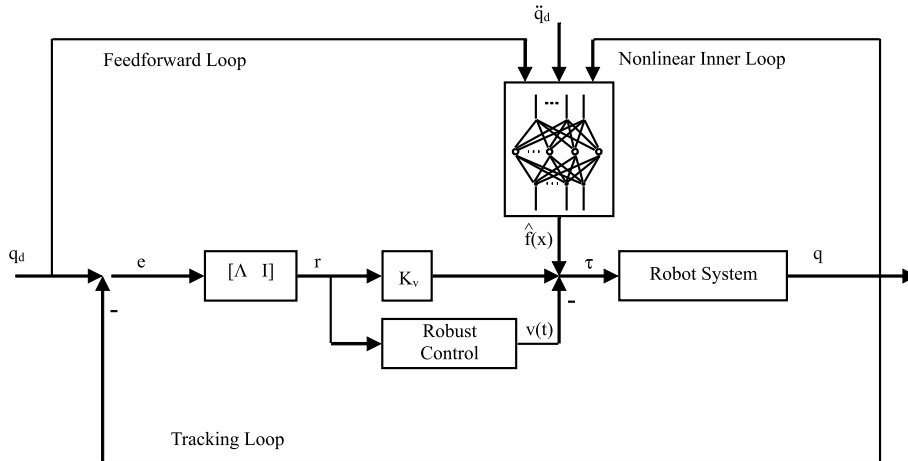
$$Z \equiv \begin{bmatrix} W & 0 \\ 0 & V \end{bmatrix} \quad (40)$$

and  $Z_B$  is a known bound for the ideal NN weights such that  $\|Z\|_F \leq Z_B$ .

The NN weights and thresholds are tuned using the following algorithms:

$$\dot{\hat{W}} = F \delta r^T - F \delta' \hat{V}^T x r^T - k_F \|r\| \hat{W} \quad (41)$$

$$\dot{\hat{V}} = G x (\delta'^T \hat{W} r)^T - k_G \|r\| \hat{V}, \quad (42)$$



Neuro-fuzzy Control of Autonomous Robotics, Figure 16  
Multilayer neural network controller structure

where the design parameters are the positive definite matrices  $F$ ,  $G$  and a small scalar  $k > 0$ . Weight initialization is not critical and there is no preliminary off-line learning required. In fact, selecting the initial weights  $\hat{W}(0)$ ,  $\hat{V}(0)$  as zero takes the NN out of the circuit and leaves only the outer PD tracking loop. It is well known that the PD term  $K_v r$  can then stabilize the plant on an interim basis until the NN begins to learn. The NN weights are tuned on-line in real time. As the NN learns, the tracking error decreases. The stability proof is given in [18]. The NN structure is shown in Fig. 16.

### Combining and Coordinating Behaviors

Typically, fuzzy controllers consider a single goal. Isolated reactive behaviors are incapable of performing autonomous navigation in complex environments. However, more complex tasks can be accomplished through combination and cooperation of primitive behaviors. For the situations where there are two or more goals active simultaneously, there are two solutions:

- Build complex rules whose antecedents consider both goals in the same time.
- Write two sets of simple rules, each set specific to a single goal, and combine their output using some form of behavior arbitration or command fusion.

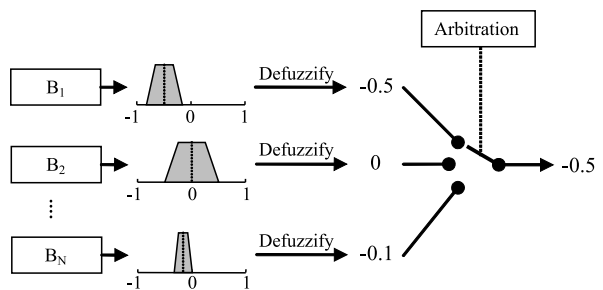
### Behavior Arbitration

Arbitration is the process that leads to the activation of a specific behavior or that generates weights for multiple simultaneous behaviors. It decides which behaviors should influence the operation of the robot and how much.

The subsumption architecture implemented by Brooks in [3] is representative for fixed behavior arbitration. Because the robot was doing a single task, he used a fixed arbitration policy, built as a network of suppression and inhibition links. However, autonomous robots need to adapt to the environment and to perform multiple tasks. As a result, the arbitration strategies have to be dynamic, taking into consideration both the environment configuration and the current plan from the higher decisional levels. By using fuzzy logic to implement either fixed or dynamic arbitration, the transition between behaviors will be smooth and it will be possible to allow partial and concurrent activation of behaviors.

### Command Fusion

The most simple and obvious method used to fuse the commands coming from multiple behaviors consists in a switching scheme (Fig. 17). The behaviors are activated only one at a time. The selected dominant behav-



Neuro-fuzzy Control of Autonomous Robotics, Figure 17  
Switching scheme for command fusion

ior solely controls the robot until the next decision cycle, whereas the motor commands of the suppressed behaviors are completely ignored. The arbitration strategy determines which behavior is activated. This switching approach achieves a poor performance in the presence of multiple goals. A good example is the potential field navigation using a reactive behavior for the avoidance of obstacles that are not represented on the map. There are two goals: the global behavior tries to reach the target, while the reactive behavior has to avoid the obstacles. When the robot approaches an obstacle, the control is taken away from the target-reaching behavior and given to the obstacle avoidance behavior. The latter can take the arbitrary decision of passing the obstacle through the left or through the right. Even if the goal of avoiding the obstacle is reached, the global solution to the navigation problem can suffer significantly if the wrong direction is chosen.

Parallel execution of multiple behaviors can overcome some of the limitations of the switching scheme. The final commands given to the robot can be calculated in two ways: by combining individual decisions (Fig. 18) or by combining individual preferences (Fig. 19).

The vector summation scheme is a good example of combining individual decisions. Considering the previous problem relating to the potential field navigation, we can identify two behaviors. Each one provides a vector characterizing the desired velocity and direction of movement. By summing the two (weighted) vectors, it is possible to obtain an intermediate speed and direction in order to

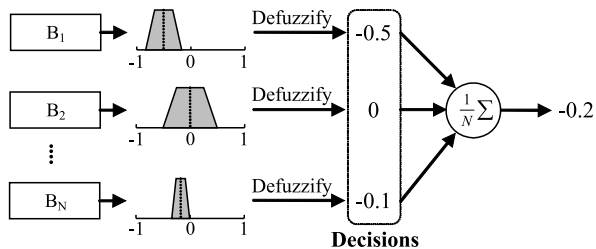
reach both goals. This approach does not necessarily need fuzzy logic, although it can be easily implemented using weighted singletons as fuzzy outputs and center-of-gravity defuzzification.

The real advantage of using fuzzy logic for command fusion becomes evident when combining individual preferences. A preference can be represented as a probability density function or as a fuzzy set [23]. It provides more information than a decision, which is a single (crisp) value. It gives information about an entire range of possible values and about how desirable they are for accomplishing the required task. Fuzzy logic has many different operators to perform combination and many defuzzification functions to perform decision. If the behaviors are also implemented using fuzzy logic, it is easy to have them output a fuzzy set instead of a crisp value, by eliminating the defuzzification phase. Even behaviors implemented using a different mechanism, but that output a PDF, can be easily accommodated.

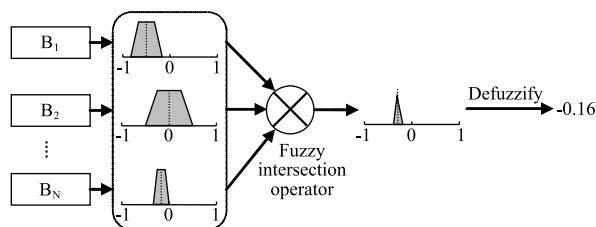
All the previous methods of command fusion suffer in some way or another when they have to handle competing behaviors that issue conflicting control commands. In this case, the resulting motor command from the compromise decision might be sub-optimal or even worse than any of the individual commands. An example could be the situation where the defuzzification results in the selection of a value that lies between two peaks of the combined fuzzy set. For a robot that has to avoid an obstacle by going to the left or to the right, this would make it go forward, straight into it. There is a need for extensions to basic fuzzy command fusion schemes capable of resolving conflicts among contradicting actions. A simple example is given in [35], where Yen proposes a different defuzzification approach, by replacing the center of gravity with centroid of largest defuzzification. This only considers the output fuzzy set with the largest area and completely ignores all the others, making the method similar to the majority voting schemes.

Context-dependent blending of behaviors is the most general type of behavior combination that can be realized using fuzzy logic. The method was suggested by Ruspini in [24] for the Flakey robot and later by Saffiotti in [26]. They reintroduce some form of behavior arbitration into fusion by having a set of higher-level supervisory fuzzy rules to activate and deactivate the individual fuzzy behaviors. As a consequence, the hierarchical behavior architecture is composed of two distinct layers. On the higher level, behavior coordination is achieved by means of supervisory fuzzy rules of the form:

IF *context* THEN *behavior*

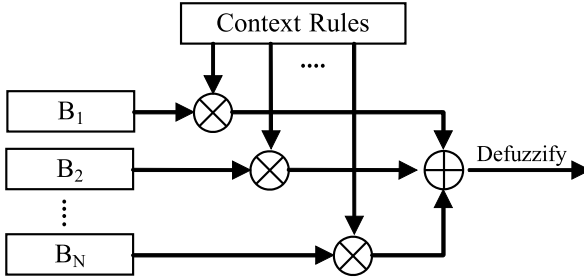


Neuro-fuzzy Control of Autonomous Robotics, Figure 18  
Command fusion by combining individual decisions



Neuro-fuzzy Control of Autonomous Robotics, Figure 19  
Command fusion by combining individual preferences





Neuro-fuzzy Control of Autonomous Robotics, Figure 20  
Context-dependent blending of behaviors [27]

As seen in Fig. 20, each behavior generates preferences in order to reach its goal. There is a context of activation for each behavior. It describes the applicability and desirability of that particular behavior and also reflects the needs of higher-level goals. The preferences of all behaviors, weighted by the truth value of their contexts, are combined to obtain the collective preference. The crisp value for the associated command is obtained after defuzzification.

The decisions can be event-driven or goal-driven. For example, in the potential field navigation problem with a reactive behavior for unknown obstacle avoidance, the following event-driven rules can apply:

IF *obstacle-close* THEN *avoid-obstacle*  
IF *not(obstacle-close)* THEN *go-to-target*

Depending on the value of the *obstacle-close* fuzzy variable, the behaviors are activated with various strengths: for extreme values, either *avoid-obstacle* or *go-to-target* is fully enabled; for intermediate values, both behaviors are partially enabled. The goal-driven approach can be used, for example, to sequence behaviors. If the robot has to pass through a set of waypoints, these rules can be used:

IF  $W_1$  *not reached* THEN *go-to- $W_1$*   
IF  $W_1$  *reached* AND  $W_2$  *not reached* THEN *go-to- $W_2$*   
...  
IF  $W_{N-1}$  *reached* AND  $W_N$  *not reached* THEN  
*go-to- $W_N$*

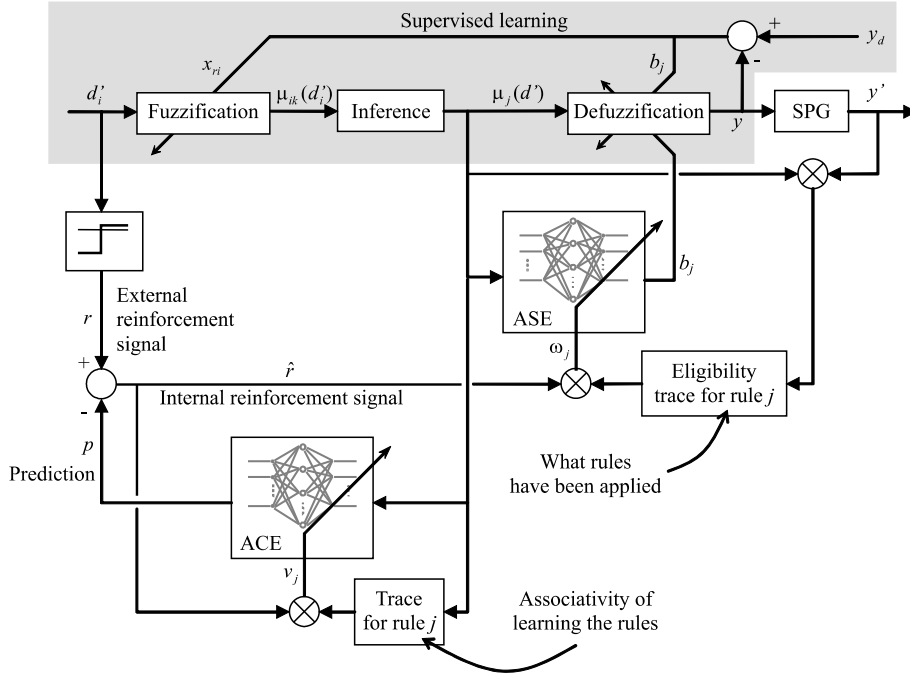
The path followed by the robot will be smooth due to fuzzy interpolation of behaviors. The direction of movement and the speed will not change suddenly when a waypoint is bypassed. The event-driven and the goal-driven blending of behaviors can be combined into an arbitrarily complex set of rules in order to represent a full plan of action.

## Learning of Behaviors

Until now we have discussed different applications of fuzzy logic for intelligent behaviors. Because fuzzy logic uses a rule base that works with linguistic variables, knowledge representation is very intuitive. A human operator can easily transpose his experience about controlling the robot into a set of rules. Unfortunately, the last step necessary for a functional fuzzy logic controller is not as easy. The relationship between the crisp values of the signals outside the controller and the fuzzy variables inside the controller is defined using membership functions that make possible the fuzzification and the defuzzification. A membership function must be defined for each value of a linguistic variable. A human expert can not define precisely the ideal shape for each of the membership functions. At best, he can only give fair approximations, based on his own way to associate numbers with words. Trial and error methods for tuning the membership functions and the rules could be avoided and an enormous amount of human efforts would be saved if the fuzzy system somehow possesses learning abilities.

A number of learning algorithms, such as the evolutionary algorithm [12], reinforcement learning [7,9], and supervised learning [14] have been proposed to construct the fuzzy system automatically. The evolutionary algorithm applied by itself results in a very long learning process. Although the reinforcing learning method seems quite promising because it doesn't require training data, it usually leads to a heavy learning phase as the gradient information is not provided explicitly and because the input space is too large for an efficient search. The supervised learning method has the advantages of fast convergence and is suitable for structure and parameter learning. However, it is difficult to obtain sufficient training data that doesn't contain conflicting input/output pairs. It is therefore impractical to learn behaviors using only one method for learning. Still, it is possible to use supervised learning to do a coarse tuning of the rule base and thus reduce the search domain for reinforcement learning, which will then fine-tune the rules. Ye, Yung and Wang apply this method in [34] for obstacle avoidance. The general aspects of their implementation will be shown here.

The neuro-fuzzy system in Fig. 21 receives information from the sensors as a  $n$ -dimensional input variable  $d'_i$ ,  $i = 1..n$  and generates a command  $y'$  to the robot. The premise membership functions are  $\mu_{ik}$ . For each input variable, the associated membership functions are parametrized using a certain number of parameters  $x_{ri}$ . The fuzzy logic system has rules. The fired strength of the



Neuro-fuzzy Control of Autonomous Robotics, Figure 21  
Neuro-fuzzy system with mixed learning algorithm

$j$ th rule is calculated as

$$\mu_j(d') = \prod_{i=1}^n \mu_{i,k_i}(d'_i) \quad (43)$$

The output membership functions are parametrized by their centers  $b_j$ .

In the first learning phase, the supervised learning method is applied. It is shown in gray in Fig. 21. For each input state vector  $\mathbf{d}$ , the system infers an output  $y$ . The difference between the desired output  $y_d$  and the system output  $y$  is used to train the fuzzy system. A steepest descent learning algorithm is used to determine the parameters of both the input ( $x_{ri}$ ) and the output ( $b_j$ ) fuzzy sets. To avoid the problem of instability and in order to increase the learning speed, a simple fuzzy system (not shown on the figure) is used to adapt the learning rate and the momentum term of the algorithm.

Reinforcement learning is done using a modified model originally proposed by Sutton and Barto [2]. For learning obstacle avoidance, there is a conflict between the desire to use the rule base already learned and the desire to further explore the environment in order to improve the rule base. This phenomenon is called “the conflict between exploration and exploitation”. To avoid using only exploitation and in order to maintain the efficiency

of learning, a stochastic perturbation generator (SPG) was added. It generates an action  $y'(t)$ , which is a Gaussian random variable with mean  $y(t)$  and a standard deviation that is high in case of a previous bad decision of the fuzzy system and small in case of a good decision. This way, the action of the SPG remains consistent with the fuzzy system when the previous action was a good one. During reinforcement learning, the parameters of the input fuzzy system are frozen and the method is applied to further tune the parameters of the output fuzzy sets.

For each input state  $d'$ , the system infers an output  $y$  that is further adjusted by adding a stochastic perturbation and finally applied to the robot. By evaluating the new state generated by the robot movement, an internal reinforcement signal is generated and used to fine-tune the fuzzy system. The process repeats at each time step until a collision occurs. At this moment a reinforcement signal  $r$ , representing failure, is fed back to the learning network and the rules used at the previous time steps are changed in order to improve the performance of the robot. This task is accomplished by an adaptive neuron-like element, which consists of an associative search element (ASE) and an associative critic element (ACE). The ACE receives an external reinforcement signal  $r(t)$  as a performance feedback and generates an internal reinforcement signal  $\hat{r}(t)$  which is fed into the ASE to update its weights. The role of the

ACE is to make predictions. In order to do so, the weights  $v_j$  of the ACE are learned through the trace of the fired rules and its own output. The weights  $\omega_j$  of the ASE are learned using the eligibility trace of each of the rules. These traces act like a filter and describe what rules have been used recently and what control actions have been applied.

Because the learning model adds perturbation to the control input of the robot, the resulting obstacle avoidance behavior is very robust to noisy sensor input. Moreover, from the experiments done by the authors in [34], the robot is able to perform collision-free navigation, it can achieve a path reasonably close to the shortest path and has a smooth motion.

### Future Directions

Some standard algorithms used for mobile robot navigation can be improved or replaced by using neuro-fuzzy methods. While many implementations require the use of very precise sensors, full models of the systems and complicated mathematics, it can be observed in nature that the same actions can be realized by animals or humans using less precise sensorial information and without having access to or needing a mathematical model of the system or the environment. Fuzzy logic and neural networks can be used to implement a similar behavior for machines. Their full potential in this area has not been reached yet.

An example of how fuzzy logic can improve a standard method heavily used in robot navigation is the implementation of a 3D dead-reckoning system with fuzzy logic instead of Kalman filtering [21]. It uses physical reasoning about sensors' strengths and weaknesses to formulate expert system rules that fuse sensor data. Many error mechanisms can be defined more accurately by expert reasoning than by the statistics-based Kalman filter methods.

### Bibliography

#### Primary Literature

- Andrews JR, Hogan N (1983) Impedance Control as a Framework for Implementing Obstacle Avoidance in a Manipulator. In: Hardt DE, Book W (eds) *Control of Manufacturing Processes and Robotic Systems*. ASME, Boston, pp 243–251
- Beom HR, Cho HS (1995) A sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning. *IEEE Trans Syst Man Cybern* 25:464–477
- Brooks RA (1986) A Robust Layered Control System For A Mobile Robot. *IEEE J Robot Autom* RA-2:14–23
- Burke RE (2001) Spinal Reflexes. In: Levitan I (ed) *Encyclopedia of Life Sciences*. Wiley, Chichester
- Chatterjee A, Matsuno F (2006) Improving EKF-based solutions for SLAM problems in Mobile Robots employing Neuro-Fuzzy Supervision. 3rd International IEEE Conference Intelligent Systems, London, pp 683–689
- Clerc M, Kennedy J (2002) The particle swarm-explosion, stability and convergence in a multidimensional complex space. *IEEE Trans Evol Comput* 6(1):58–73
- Colombetti M, Dorigo M (1996) Behavior analysis and training-A methodology for behavior engineering. *IEEE Trans Syst Man Cybern B* 26:365–380
- Dissanayake MWMG, Newman P, Clark S, Durrant-Whyte HF (2001) A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Trans Robot Autom* 17(3):229–241
- Donnart J, Meyer J (1996) Learning reactive and planning rules in a motivationally autonomous animation. *IEEE Trans Syst Man Cybern B* 26:381–395
- Donoghue JP, Sanes JN (2001) Motor System Organization. In: Levitan I (ed) *Encyclopedia of Life Sciences*. Wiley, Chichester
- Fitzgerald RJ (1971) Divergence of the Kalman filter. *IEEE Trans Autom Control* AC-16(6) 736–747
- Homaifar A, McCormick E (1995) Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. *IEEE Trans Fuzzy Syst* 3:129–139
- Hebert T (1998) Navigation of an autonomous vehicle using a combined electrostatic potential field/fuzzy inference approach. Ph.D. dissertation, University of Southwestern Louisiana, Lafayette
- Juang CF, Lin CT (1998) An on-line self-constructing neural fuzzy inference network and its applications. *IEEE Trans Fuzzy Syst* 6:12–32
- Khatib O (1985) Real-time obstacle avoidance for manipulators and mobile robots. *IEEE Int Conf Robotics and Automation*, St. Louis, MO, pp 500–505
- Lefebvre DR, Saridis GN (1991) A Computer Architecture for Intelligent Machines. *Intelligent Robotic Systems for Space Exploration*. Rensselaer Polytechnic Institute Troy, New York, pp 31–43
- Lewis FL (1986) *Optimal Estimation: With an Introduction to Stochastic Control Theory*. Wiley-Interscience, New York
- Lewis FL, Jagannathan S, Yesildirek A (1998) *Neural Network Control of Robot Manipulators and Nonlinear Systems*. Taylor & Francis, Inc., Bristol
- Loebis D, Sutton R, Chudley J, Naeem W (2004) Adaptive tuning of a Kalman filter via fuzzy logic for an intelligent AUV navigation system. *Control Eng Pract* 12:1531–1539
- Mehra RK (1970) On the identification of variances and adaptive Kalman filtering. *IEEE Trans Autom Control* AC-15(2):175–184
- Ojeda L, Borenstein J (2002) FLEXnav: Fuzzy Logic Expert Rule-based Position Estimation for Mobile Robots on Rugged Terrain. *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, Washington DC, pp 317–322
- Park J, Sandberg IW (1991) Universal approximation using radial-basis-function networks. *Neural Comput* 3:246–257
- Ruspini EH (1991) Truth as utility: a conceptual synthesis. In: *Proc 7th Conf Uncertainty in Artificial Intelligence*, Los Angeles, CA
- Ruspini EH (1990) Fuzzy logic in the Flakey robot. *Proc Int Conf Fuzzy Logic and Neural Networks (IIZUKA)*, Iizuka, Japan, pp 767–770
- Saffiotti A (1997) The Uses of Fuzzy Logic in Autonomous Robotics: a catalogue raisonne. In: Degli Antoni G (ed) *Soft Computing I(4)*. Springer, Berlin, pp 180–197

26. Saffiotti A, Konolige K, Ruspini EH (1995) A multivalued-logic approach to integrating planning and control. *Artif Intell* 76:(1–2):481–526
27. Saffiotti A, Ruspini EH, Konolige K (1993) Blending reactivity and goal-directedness in a fuzzy controller. In: *Proceedings of the IEEE Int Conf Fuzzy Systems*, San Francisco, California, pp 134–139
28. Saridis GN, Stephanou HE (1977) A hierarchical approach to the control of a prosthetic arm. *IEEE Trans Syst Man Cybern SMC*-7(6):407–420
29. Thrun S, Burgard W, Fox D (2005) *Probabilistic Robotics*. MIT Press, Cambridge
30. Valavanis KP, Hebert T, Kolluru R, Tsourveloudis NC (2000) Mobile robot navigation in 2-D dynamic environments using electrostatic potential fields. *IEEE Trans Syst Man Cybern A* 30:187–197
31. Welch G, Bishop G (2004) An introduction to the Kalman filter. Technical Report TR 95-041, University of North Carolina, Department of Computer Science, Chapel Hill
32. Wang LX (1992) Fuzzy systems are universal approximators. In: *Proceedings of the 1st IEEE Conference on Fuzzy Systems*, San Diego, pp 1163–1170
33. Werbos PJ (1998) Backpropagation: basics and new developments. *The handbook of brain theory and neural networks*. MIT Press, Cambridge, pp 134–139
34. Ye C, Yung NHC, Wang D (2003) A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance. *IEEE Trans Syst Man Cybern B* 33(1):17–27
35. Yen J, Pfluger N (1995) A Fuzzy Logic Based Extension to Payton and Rosenblatt's Command Fusion Method for Mobile Robot Navigation. *IEEE Trans Syst Man Cybern* 25(6):971–978
36. Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans Syst Man Cybern SMC*-3:28–44

## Books and Reviews

- Ge SS, Lewis FL (2006) *Autonomous Mobile Robots: Sensing, Control, Decision Making and Applications*. *Autom Control Eng Ser* 22:229–265
- Tzafestas SG (1999) *Advances in Intelligent Autonomous Systems*. Microprocessor-Based and Intelligent Systems Engineering Series, vol 18. Springer, New York
- Bekey GA (2005) *Autonomous Robots: From Biological Inspiration to Implementation and Control*. *Intelligent Robotics and Autonomous Agents Series*, MIT Press, Cambridge

## Neuro-fuzzy Systems

LESZEK RUTKOWSKI<sup>1</sup>, KRZYSZTOF CPAŁKA<sup>1,2</sup>,  
ROBERT NOWICKI<sup>1</sup>, AGATA POKROPIŃSKA<sup>3</sup>,  
RAFAŁ SCHERER<sup>1</sup>

<sup>1</sup> Department of Computer Engineering, Częstochowa University of Technology, Częstochowa, Poland

<sup>2</sup> Department of Artificial Intelligence, Academy of Humanities and Economics, Lodz, Poland

<sup>3</sup> Institute of Mathematics and Computer Science, Jan Długosz University, Częstochowa, Poland

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Fuzzy Reasoning](#)

[Description of Fuzzy Inference Systems](#)

[Logical-Type Neuro-fuzzy Systems](#)

[Mamdani-Type Neuro-fuzzy Systems](#)

[Simplified Neuro-fuzzy Systems](#)

[Takagi–Sugeno Neuro-fuzzy Systems](#)

[Neuro-fuzzy Systems with Weights](#)

[Neuro-fuzzy Systems for Pattern Classification](#)

[Learning](#)

[Criteria Isolines Method](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

## Glossary

**Neuro-fuzzy systems** Fusion of fuzzy logic and neural networks with the ability to automated adaptation to training data and knowledge interpretability.

**Fuzzy reasoning** Reasoning on the basis of fuzzy premises and fuzzy rules inferring fuzzy conclusions.

## Definition of the Subject

Neuro-fuzzy systems (NFS) are part of soft computing concept. They are a synergistic fusion of fuzzy logic and neural networks with the ability to automate adaptation to training data and knowledge interpretability. Thus neuro-fuzzy systems incorporate two goals of machine learning which are difficult to meet at the same time: transparency of the rule base and high accuracy obtained thanks to learning from data. There are various types of NFS which differ in the form of fuzzy rules and inference methods. They have a very broad range of applications in life, science, economics, manufacturing, medicine etc. NFS can replace neural networks in all applications as most of them are universal approximators. Therefore they can perform well in tasks of classification, prediction, approximation and control, similarly to neural networks. Neural networks work as black boxes (usually it is not possible to tell where the knowledge is stored) and can not use prior knowledge. NFS can use almost the same learning methods and achieve the same accuracy as neural networks yet the knowledge in the form of fuzzy rules is easily interpretable for humans. On the other hand, fuzzy systems (not neuro-fuzzy) require manual tuning of their parameters and fuzzy reasoning in such systems is more trou-

blesome and computationally demanding than simplified one, used in NFS.

## Introduction

Over the last decade fuzzy sets and fuzzy logic introduced in 1965 by Lotfi Zadeh [62] have been used in a wide range of problem domains, including process control, image processing, pattern recognition and classification, management, economics and decision making. Specific applications include washing-machine automation, camcorder focusing, TV color tuning, automobile transmissions and subway operations [17]. We have also been witnessing a rapid development in the area of neural networks (see e.g. [53,54,63]). Both fuzzy systems and neural networks, along with probabilistic methods [1,11,42], evolutionary algorithms [13,36], rough sets [43,44] and uncertain variables [3,4,5], constitute a consortium of soft computing techniques [1,23,26]. These techniques are often used in combination. For example, fuzzy inference systems are frequently converted into connectionist structures called neuro-fuzzy systems which exhibit advantages of neural networks and fuzzy systems. In literature various neuro-fuzzy systems have been developed (see e.g. [6,7,8,9,10,14,15,16,19,20,21,23,28,29,30,31,32,33,34,37,38,41,45,46,47,49,50,58]). They combine the natural language description of fuzzy systems and the learning properties of neural networks. Some of them are known in literature under short names such as ANFIS [18], ANNBFIS [9], DENFIS [24], FALCON [31], GARIC [2], NEFCLASS [40], NEFPROX [39,40], SANFIS [57] and others. The most popular designs of neuro-fuzzy structures fall into one of the following categories, depending on the connective between the antecedent and the consequent in fuzzy rules:

- (i) Takagi–Sugeno method – consequents are functions of inputs,
- (ii) Mamdani-type reasoning method – consequents and antecedents are related by the min operator or generally by a t-norm,
- (iii) Logical-type reasoning method – consequents and antecedents are related by fuzzy implications, e. g. binary, Łukasiewicz, Zadeh and others.

It should be noted that most applications are dominated by the Mamdani-type fuzzy reasoning. However, it was emphasized by Yager [60,61] that “no formal reason exists for the preponderant use of the Mamdani method in fuzzy logic control as opposed to the logical method other than inertia”. In this work we will outline all three types of sys-

tems. Moreover we will compare several variants of these systems in terms of efficiency for an exemplary problem.

## Fuzzy Reasoning

In this section we present the idea of fuzzy reasoning with various fuzzy implications which will be useful for construction of neuro-fuzzy systems developed in the next sections. The basic rule of inference in classical logic is modus ponens. The compositional rule of inference describes a composition of a fuzzy set and a fuzzy relation. Fuzzy rule

$$\text{IF } x \text{ is } A \text{ THEN } y \text{ is } B \quad (1)$$

is represented by a fuzzy relation  $R$ . Having given input linguistic value  $A'$ , we can infer an output fuzzy set  $B'$  by the composition of the fuzzy set  $A'$  and the relation  $R$ . The generalized modus ponens is the extension of the conventional modus ponens tautology, to allow partial fulfillment of the premises:

Premise	$x \text{ is } A'$
Implication	IF $x \text{ is } A$ THEN $y \text{ is } B$
Conclusion	$y \text{ is } B'$

where  $A, B, A', B'$  are fuzzy sets,  $x$  and  $y$  are linguistic variables. Applying the compositional rule of inference [22], we obtain

$$B' = A' \circ R = A' \circ (A \rightarrow B) \quad (2)$$

and

$$\mu_{B'}(y) = \mu_{A' \circ R}(y) = \sup_{x \in X} \{ \mu_{A'}(x)^T * \mu_R(x, y) \} \quad (3)$$

The problem is to determine the membership function of the fuzzy relation described by

$$\mu_R(x, y) = \mu_{A \rightarrow B}(x, y) \quad (4)$$

based on the knowledge of  $\mu_A(x)$  and  $\mu_B(y)$ . We denote

$$\mu_{A \rightarrow B}(x, y) = I(\mu_A(x), \mu_B(y)), \quad (5)$$

where  $I(\cdot) = I_{\text{fuzzy}}(\cdot)$  is a fuzzy implication (logical approach to fuzzy reasoning, see Subsect. “[Logical Approach to Fuzzy Inference Reasoning](#)”) given in Definition 1 or a t-norm (Mamdani approach to fuzzy reasoning, see Subsect. “[Mamdani Approach to Fuzzy Inference Reasoning](#)”).



**Neuro-fuzzy Systems, Table 1**  
**Fuzzy implications**

No	Name	Implication $I_{\text{fuzzy}}(a, b)$
1	Kleene–Dienes (binary)	$\max\{1 - a, b\}$
2	Łukasiewicz	$\min\{1, 1 - a + b\}$
3	Reichenbach	$1 - a + a \cdot b$
4	Fodor	$\begin{cases} 1 & \text{if } a \leq b \\ \max\{1 - a, b\} & \text{if } a > b \end{cases}$
5	Rescher	$\begin{cases} 1 & \text{if } a \leq b \\ 0 & \text{if } a > b \end{cases}$
6	Goguen	$\begin{cases} 1 & \text{if } a = 0 \\ \min\{1, \frac{b}{a}\} & \text{if } a > 0 \end{cases}$
7	Gödel	$\begin{cases} 1 & \text{if } a \leq b \\ b & \text{if } a > b \end{cases}$
8	Yager	$\begin{cases} 1 & \text{if } a = 0 \\ b^a & \text{if } a > 0 \end{cases}$
9	Zadeh	$\max\{\min\{a, b\}, 1 - a\}$
10	Willmott	$\min\left\{\begin{matrix} \max\{1 - a, b\}, \\ \max\{a, 1 - b, \min\{1 - a, b\}\} \end{matrix}\right\}$
11	Dubois–Prade	$\begin{cases} 1 - a & \text{if } b = 0 \\ b & \text{if } a = 1 \\ 1 & \text{if otherwise} \end{cases}$

### Logical Approach to Fuzzy Inference Reasoning

**Definition 1** (see Fodor [12]) A fuzzy implication is a function  $I = I_{\text{fuzzy}}: [0, 1]^2 \rightarrow [0, 1]$  satisfying the following conditions:

- (I1) if  $a_1 \leq a_3$ , then  $I(a_1, a_2) \geq I(a_3, a_2)$ , for all  $a_1, a_2, a_3 \in [0, 1]$ ,
- (I2) if  $a_2 \leq a_3$ , then  $I(a_1, a_2) \leq I(a_1, a_3)$ , for all  $a_1, a_2, a_3 \in [0, 1]$ ,
- (I3)  $I(0, a_2) = 1$ , for all  $a_2 \in [0, 1]$  (falsity implies anything),
- (I4)  $I(a_1, 1) = 1$ , for all  $a_1 \in [0, 1]$  (anything implies tautology),
- (I5)  $I(1, 0) = 0$  (Booleanity).

Selected fuzzy implications satisfying all or some of the above conditions are listed in Table 1.

In this table, Implications 1–4 are examples of an S-implication associated with a t-conorm

$$I(a, b) = S\{1 - a, b\}. \quad (6)$$

Neuro-fuzzy systems based on fuzzy implications given in Table 1 are called logical systems. Implications (6) and (7) belong to a group of R-implications associated with the t-norm  $T$  and given by

$$I(a, b) = \sup_z \{z | T\{a, z\} \leq b\}, \quad a, b \in [0, 1]. \quad (7)$$

The Zadeh implication belongs to a group of Q-implications given by

$$I(a, b) = S\{N(a), T\{a, b\}\}, \quad a, b \in [0, 1]. \quad (8)$$

It is easy to verify that S-implications and R-implications satisfy all the conditions of Definition 1. However, the Zadeh implication violates conditions I1 and I4, whereas the Willmott implication violates conditions I1, I3 and I4.

### Mamdani Approach to Fuzzy Inference Reasoning

In practice we frequently use Mamdani-type operators given by

$$I(a, b) = \min\{a, b\}, \quad a, b \in [0, 1], \quad (9)$$

$$I(a, b) = a \cdot b, \quad a, b \in [0, 1] \quad (10)$$

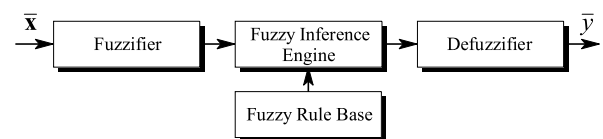
or generally

$$I(a, b) = T\{a, b\}, \quad a, b \in [0, 1]. \quad (11)$$

It should be noted that operators (9)–(11) do not satisfy conditions of Definition 1. Operators (9)–(11) are called “correlation functions” describing strength of connections between antecedents and consequences (see [48]).

### Description of Fuzzy Inference Systems

In this description, we consider multi-input-single-output fuzzy system mapping  $\mathbf{X} \rightarrow \mathbf{Y}$ , where  $\mathbf{X} \subset \mathbf{R}^n$  and  $\mathbf{Y} \subset \mathbf{R}$  (see Fig. 1). The system is composed of a fuzzifier, a fuzzy rule base, a fuzzy inference engine and a defuzzifier. The fuzzifier performs a mapping from the observed crisp input space  $\mathbf{X} \subset \mathbf{R}^n$  to a fuzzy set defined in  $\mathbf{X}$ . The most commonly used fuzzifier is the singleton fuzzifier, which



**Neuro-fuzzy Systems, Figure 1**  
**Fuzzy inference system**

maps  $\bar{\mathbf{x}} = [\bar{x}_1, \dots, \bar{x}_n] \in \mathbf{X}$  into a fuzzy set  $A' \subseteq \mathbf{X}$  characterized by the membership function

$$\mu_{A'}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} = \bar{\mathbf{x}} \\ 0 & \text{if } \mathbf{x} \neq \bar{\mathbf{x}} \end{cases} \quad (12)$$

The fuzzy rule base consists of a collection of  $N$  fuzzy IF-THEN rules, aggregated by the disjunction or the conjunction, in the form

$$R^{(k)} : \begin{cases} \text{IF} & x_1 \text{ is } A_1^k \text{ AND} \\ & x_2 \text{ is } A_2^k \text{ AND} \dots \\ & x_n \text{ is } A_n^k \\ \text{THEN} & y \text{ is } B^k \end{cases} \quad (13)$$

or

$$R^{(k)} : \text{IF } \mathbf{x} \text{ is } A^k \text{ THEN } y \text{ is } B^k, \quad (14)$$

where  $\mathbf{x} = [x_1, \dots, x_n] \in \mathbf{X}$ ,  $y \in \mathbf{Y}$ ,  $A^k = A_1^k \times A_2^k \times \dots \times A_n^k$ ,  $A_1^k, A_2^k, \dots, A_n^k$  are fuzzy sets characterized by membership functions  $\mu_{A_i^k}(x_i)$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, N$ , whereas  $B^k$  are fuzzy sets characterized by membership functions  $\mu_{B^k}(y)$ ,  $k = 1, \dots, N$ . The firing strength of the  $k$ th rule,  $k = 1, \dots, N$ , is defined by

$$\tau_k(\bar{\mathbf{x}}) = \bigwedge_{i=1}^n \{\mu_{A_i^k}(\bar{x}_i)\} = \mu_{A^k}(\bar{\mathbf{x}}). \quad (15)$$

The fuzzy inference engine determines the mapping from the fuzzy sets in the input space  $\mathbf{X}$  to the fuzzy sets in the output space  $\mathbf{Y}$ . Each of  $N$  rules (14) determines a fuzzy set  $\bar{B}^k \subseteq \mathbf{Y}$  given by the compositional rule of inference

$$\bar{B}^k = A' \circ (A^k \rightarrow B^k), \quad (16)$$

where  $A^k = A_1^k \times A_2^k \times \dots \times A_n^k$ . Fuzzy sets  $\bar{B}^k$  are characterized by membership functions expressed by the sup-star composition

$$\mu_{\bar{B}^k}(y) = \sup_{\mathbf{x} \in \mathbf{X}} \{\mu_{A'}(\mathbf{x}) * \mu_{A_1^k \times \dots \times A_n^k \rightarrow B^k}(\mathbf{x}, y)\}, \quad (17)$$

where  $*$  can be any operator in the class of t-norms. It is easily seen that for a crisp input  $\bar{\mathbf{x}} = [\bar{x}_1, \dots, \bar{x}_n] \in \mathbf{X}$ , i. e., the singleton fuzzifier (12), formula (17) becomes

$$\begin{aligned} \mu_{\bar{B}^k}(y) &= \mu_{A_1^k \times \dots \times A_n^k \rightarrow B^k}(\bar{\mathbf{x}}, y) \\ &= \mu_{A^k \rightarrow B^k}(\bar{\mathbf{x}}, y) \\ &= I(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(y)), \end{aligned} \quad (18)$$

where  $I(\cdot)$  is correlation between antecedents and consequents given by (11), in the case of Mamdani approach, or

fuzzy implication  $I_{\text{fuzzy}}(\cdot)$ , in the case of logical approach (examples of fuzzy implications are given in Table 1). The aggregation operator, applied in order to obtain the fuzzy set  $B'$  based on fuzzy sets  $\bar{B}^k$ , is the t-norm or t-conorm operator, depending on the type of the fuzzy inference. As a result of the fuzzy reasoning we obtain the fuzzy set  $B'$ .

The defuzzifier performs a mapping from the fuzzy set  $B'$  to a crisp point  $\bar{y}$  in  $\mathbf{Y} \subseteq \mathbf{R}$ . The COA (center of area) method is defined by the following formula:

$$\bar{y} = \frac{\int_{\mathbf{Y}} y \cdot \mu_{B'}(y) dy}{\int_{\mathbf{Y}} \mu_{B'}(y) dy} \quad (19)$$

or by

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \cdot \mu_{B'}(\bar{y}^r)}{\sum_{r=1}^N \mu_{B'}(\bar{y}^r)}, \quad (20)$$

in the discrete form, where  $\bar{y}^r$  are centers of the membership functions  $\mu_{B'}(y)$ , i. e., for  $r = 1, \dots, N$

$$\mu_{B'}(\bar{y}^r) = \max_{y \in \mathbf{Y}} \{\mu_{B^k}(y)\}. \quad (21)$$

### Logical-Type Neuro-fuzzy Systems

In this approach, function  $I(\cdot)$  given by 18 is a fuzzy implication (see Table 1), i. e.

$$I(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r)) = I_{\text{fuzzy}}(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r)). \quad (22)$$

When we use the logical model, the aggregation is carried out by

$$B' = \bigcap_{k=1}^N \bar{B}^k. \quad (23)$$

The aggregated output fuzzy set  $B' \subseteq \mathbf{Y}$  is computed by the use of a t-norm and is given by

$$\mu_{B'}(\bar{y}^r) = \bigwedge_{k=1}^N \{\mu_{\bar{B}^k}(\bar{y}^r)\} = \bigwedge_{k=1}^N \{I_{\text{fuzzy}}(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r))\} \quad (24)$$

and formula (20) becomes

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \cdot \bigwedge_{k=1}^N \left\{ I_{\text{fuzzy}} \left( \bigwedge_{i=1}^n \{\mu_{A_i^k}(\bar{x}_i)\}, \mu_{B^k}(\bar{y}^r) \right) \right\}}{\sum_{r=1}^N \bigwedge_{k=1}^N \left\{ I_{\text{fuzzy}} \left( \bigwedge_{i=1}^n \{\mu_{A_i^k}(\bar{x}_i)\}, \mu_{B^k}(\bar{y}^r) \right) \right\}}. \quad (25)$$

*Example 1* Using Gaussian membership functions

$$\mu_{A_i^r}(x_i) = \exp \left[ - \left( \frac{x_i - \tilde{x}_i^r}{\sigma_i^r} \right)^2 \right], \quad (26)$$

$$\mu_{B^r}(y) = \exp \left[ - \left( \frac{y - \tilde{y}^r}{\sigma^r} \right)^2 \right], \quad (27)$$

dependency (25), Łukasiewicz-type S-implication (see Table 1), and the min-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system

$$\bar{y} = \frac{\sum_{r=1}^N \tilde{y}^r \cdot \min_{k=1, \dots, N} \left\{ \min \left[ 1, 1 - \frac{n}{T} \left( \exp \left[ - \left( \frac{\tilde{x}_i - \tilde{x}_i^k}{\sigma_i^k} \right)^2 \right] \right) + \exp \left[ - \left( \frac{\tilde{y}^r - \tilde{y}^k}{\sigma^k} \right)^2 \right] \right] \right\}}{\sum_{r=1}^N \cdot \min_{k=1, \dots, N} \left\{ \min \left[ 1, 1 - \frac{n}{T} \left( \exp \left[ - \left( \frac{\tilde{x}_i - \tilde{x}_i^k}{\sigma_i^k} \right)^2 \right] \right) + \exp \left[ - \left( \frac{\tilde{y}^r - \tilde{y}^k}{\sigma^k} \right)^2 \right] \right] \right\}} \quad (28)$$

*Example 2* Using Gaussian membership functions (26) and (27), dependency (25), Kleene–Dienes-type S-implications (see Table 1), and the min-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system:

$$\bar{y} = \frac{\sum_{r=1}^N \tilde{y}^r \cdot \min_{k=1, \dots, N} \left\{ \max \left[ 1 - \frac{n}{T} \left( \exp \left[ - \left( \frac{\tilde{x}_i - \tilde{x}_i^k}{\sigma_i^k} \right)^2 \right] \right), \exp \left[ - \left( \frac{\tilde{y}^r - \tilde{y}^k}{\sigma^k} \right)^2 \right] \right] \right\}}{\sum_{r=1}^N \cdot \min_{k=1, \dots, N} \left\{ \max \left[ 1 - \frac{n}{T} \left( \exp \left[ - \left( \frac{\tilde{x}_i - \tilde{x}_i^k}{\sigma_i^k} \right)^2 \right] \right), \exp \left[ - \left( \frac{\tilde{y}^r - \tilde{y}^k}{\sigma^k} \right)^2 \right] \right] \right\}} \quad (29)$$

*Example 3* Using Gaussian membership functions (26) and (27), dependency (25), Reichenbach-type S-implications (see Table 1), and the min-type aggregation of con-

clusions, we obtain the following description of the neuro-fuzzy system:

$$\bar{y} = \frac{\sum_{r=1}^N \tilde{y}^r \cdot \min_{k=1, \dots, N} \left\{ 1 - \frac{n}{T} \left( \exp \left[ - \left( \frac{\tilde{x}_i - \tilde{x}_i^k}{\sigma_i^k} \right)^2 \right] \right) \cdot \left( 1 - \exp \left[ - \left( \frac{\tilde{y}^r - \tilde{y}^k}{\sigma^k} \right)^2 \right] \right) \right\}}{\sum_{r=1}^N \cdot \min_{k=1, \dots, N} \left\{ 1 - \frac{n}{T} \left( \exp \left[ - \left( \frac{\tilde{x}_i - \tilde{x}_i^k}{\sigma_i^k} \right)^2 \right] \right) \cdot \left( 1 - \exp \left[ - \left( \frac{\tilde{y}^r - \tilde{y}^k}{\sigma^k} \right)^2 \right] \right) \right\}} \quad (30)$$

### Mamdani-Type Neuro-fuzzy Systems

In this approach, function  $I(\cdot)$  given by (18) is a t-norm (e. g. minimum or algebraic), i. e.

$$I(\mu_{A^k}(\tilde{\mathbf{x}}), \mu_{B^k}(\tilde{y}^r)) = T\{\mu_{A^k}(\tilde{\mathbf{x}}), \mu_{B^k}(\tilde{y}^r)\}. \quad (31)$$

In case of the Mamdani approach, the aggregation is carried out by

$$B' = \bigcup_{k=1}^N \tilde{B}^k. \quad (32)$$

The aggregated output fuzzy set  $B' \subseteq \mathbf{Y}$  is computed by the use of a t-conorm and is given by

$$\mu_{B'}(\tilde{y}^r) = \bigvee_{k=1}^N \{\mu_{\tilde{B}^k}(\tilde{y}^r)\} = \bigvee_{k=1}^N \{T\{\mu_{A^k}(\tilde{\mathbf{x}}), \mu_{B^k}(\tilde{y}^r)\}\}. \quad (33)$$

Consequently, formula (20) takes the form

$$\bar{y} = \frac{\sum_{r=1}^N \tilde{y}^r \cdot \bigvee_{k=1}^N \left\{ T \left\{ \frac{n}{T} \{\mu_{A_i^k}(\tilde{x}_i), \mu_{B^k}(\tilde{y}^r)\} \right\} \right\}}{\sum_{r=1}^N \bigvee_{k=1}^N \left\{ T \left\{ \frac{n}{T} \{\mu_{A_i^k}(\tilde{x}_i), \mu_{B^k}(\tilde{y}^r)\} \right\} \right\}} \quad (34)$$

Obviously, the t-norms used to connect the antecedents in the rules and to connect the antecedents and consequents (correlation function) do not have to be the same. Besides, they can be chosen as differentiable functions as e. g. Yager families [27].

**Remark 1** Another Mamdani-type neuro-fuzzy system can be derived using so-called “height defuzzifier”

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \cdot \mu_{\bar{B}^r}(\bar{y}^r)}{\sum_{r=1}^N \mu_{\bar{B}^r}(\bar{y}^r)}. \quad (35)$$

For normal fuzzy sets  $B^r$ , combining (35), (18) and (31), we obtain

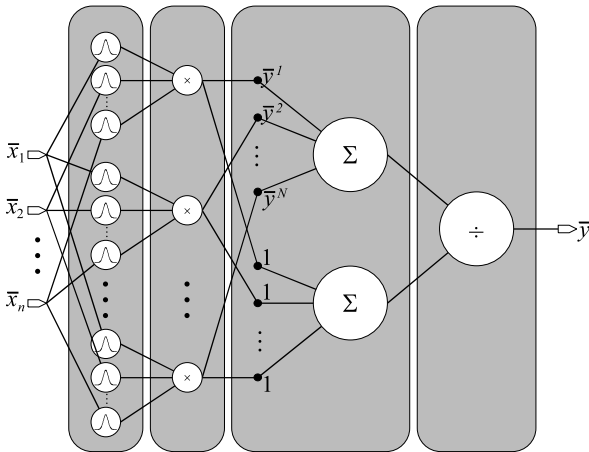
$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \cdot \frac{1}{T} \sum_{i=1}^n \{\mu_{A_i^r}(\bar{x}_i)\}}{\sum_{r=1}^N \frac{1}{T} \sum_{i=1}^n \{\mu_{A_i^r}(\bar{x}_i)\}}, \quad (36)$$

and for Gaussian functions (26) and (27) and the product t-norm we get

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \cdot \left[ \prod_{i=1}^n \left( \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r} \right)^2 \right] \right) \right]}{\sum_{r=1}^N \left[ \prod_{i=1}^n \left( \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r} \right)^2 \right] \right) \right]}. \quad (37)$$

It should be noted that (37) is a well-known formula derived by Wang [58] and used for tasks of modeling and control. It is interesting that formula (37) can be also derived from defuzzifier (20) if assumption (40) holds (see Sect. “Fuzzy Reasoning”). The architecture of the system (37) is depicted in Fig. 2.

**Example 4** Using Gaussian membership functions (26) and (27), dependency (34), min-type Mamdani rule, and



**Neuro-fuzzy Systems, Figure 2**

Architecture of the neuro-fuzzy system with Mamdani approach described by (37)

max-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \cdot \max_{k=1, \dots, N} \left\{ \min \left( \prod_{i=1}^n \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right) \cdot \exp \left[ - \left( \frac{\bar{y}^r - \bar{y}^k}{\sigma^k} \right)^2 \right] \right\}}{\sum_{r=1}^N \max_{k=1, \dots, N} \left\{ \min \left( \prod_{i=1}^n \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right) \cdot \exp \left[ - \left( \frac{\bar{y}^r - \bar{y}^k}{\sigma^k} \right)^2 \right] \right\}}. \quad (38)$$

**Example 5** Using Gaussian membership functions (26) and (27), dependency (34), product-type Mamdani rule (known as the Larsen rule), and the max-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system:

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \cdot \max_{k=1, \dots, N} \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \cdot \exp \left[ - \left( \frac{\bar{y}^r - \bar{y}^k}{\sigma^k} \right)^2 \right] \right\}}{\sum_{r=1}^N \max_{k=1, \dots, N} \left\{ \prod_{i=1}^n \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \cdot \exp \left[ - \left( \frac{\bar{y}^r - \bar{y}^k}{\sigma^k} \right)^2 \right] \right\}}. \quad (39)$$

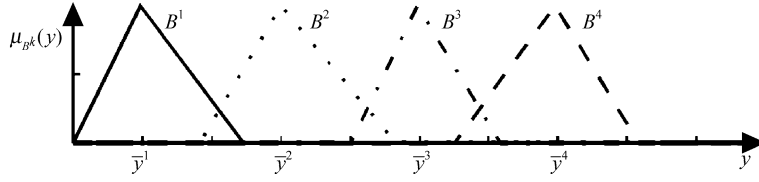
### Simplified Neuro-fuzzy Systems

The structures presented so far are quite complex and complicated. One possibility of their simplification is to assume, that membership functions of consequent fuzzy sets are sufficiently distant from each other, so that the following assumption holds

$$\mu_{B^k}(\bar{y}^r) \approx 0, \quad (40)$$

for  $k, r = 1, \dots, N$  and  $k \neq r$ . Figure 3 shows an example of such fuzzy sets.

It is easy to verify that under assumption (40) the Mamdani-type neuro-fuzzy structure given by (34), derived from defuzzifier (20), takes the simplified form



**Neuro-fuzzy Systems, Figure 3**  
Exemplary fuzzy sets satisfying assumption (40)

which is the same as the structure (36) derived from defuzzifier (35). The logical neuro-fuzzy system (25) can also be simplified if condition (40) holds, e.g. the Lukasiewicz neuro-fuzzy system (28) takes the form

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \prod_{\substack{k=1 \\ k \neq r}}^N \left\{ 1 - \frac{n}{i=1} \left\{ \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right\} \right\}}{\sum_{r=1}^N \prod_{\substack{k=1 \\ k \neq r}}^N \left\{ 1 - \frac{n}{i=1} \left\{ \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right\} \right\}}. \quad (41)$$

Let us observe that simplified neuro-fuzzy systems are characterized by three parameters to be found in the process of learning:  $\bar{x}_i^k$ ,  $\sigma_i^k$  and  $\bar{y}^k$  for  $i = 1, \dots, n$  and  $k = 1, \dots, N$ . In the case of not simplified systems, see (28)–(30), we have to additionally determine parameters  $\sigma^k$ ,  $k = 1, \dots, N$ , of the output membership functions. The architecture of the system (41) is depicted in Fig. 3.

### Takagi–Sugeno Neuro-fuzzy Systems

In the fuzzy Takagi–Sugeno-type model [55], the rules have fuzzy character only in the IF part, whereas in the THEN part, there are functional dependencies

$$\begin{aligned} R^{(r)}: & \text{ IF } \mathbf{x} \text{ is } A^r \\ & \text{ THEN } y_r = f^{(r)}(x_1, x_2, \dots, x_n). \end{aligned} \quad (42)$$

If we assume that the input of the fuzzy system is signal  $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$ , then in order to obtain the output signal  $\bar{y}$  of the system, first we will determine

$$T(\mu_{A_1^r}(\bar{x}_1), \mu_{A_2^r}(\bar{x}_2), \dots, \mu_{A_n^r}(\bar{x}_n)), \quad r = 1, \dots, N. \quad (43)$$

The next step is to compute

$$\bar{y}_r = f^{(r)}(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n), \quad r = 1, \dots, N. \quad (44)$$

The output signal of the fuzzy Takagi–Sugeno system is a normalized weighted sum of particular inputs  $\bar{y}_1, \dots,$

$\bar{y}_N$ , i. e.

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}_r \prod_{i=1}^n \{\mu_{A_i^r}(\bar{x}_i)\}}{\sum_{r=1}^N \prod_{i=1}^n \{\mu_{A_i^r}(\bar{x}_i)\}}. \quad (45)$$

In the following part of this subchapter, we will consider the Takagi–Sugeno systems with linear dependencies in consequents of the base of rules, i. e.

$$\begin{aligned} R^{(r)}: & \text{ IF } \mathbf{x} \text{ is } A^r \\ & \text{ THEN } y_r = c_0^{(r)} + c_1^{(r)}x_1 + \dots + c_n^{(r)}x_n \end{aligned} \quad (46)$$

for  $r = 1, \dots, N$ . It should be noted that if  $c_i^{(r)} = 0$ ,  $i = 1, \dots, n$ , then system (45) is reduced to a simplified Mamdani system given by formula (36), and then  $c_0^{(r)} = \bar{y}^r$ ,  $r = 1, \dots, N$ .

**Example 6** The Takagi–Sugeno system with Gaussian membership functions, linear model (46) and product t-norm, takes the following form

$$\bar{y} = \frac{\sum_{r=1}^N \left[ \prod_{i=1}^n \left( \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r} \right)^2 \right] \right) \right] \cdot (c_0^{(r)} + c_1^{(r)}x_1 + \dots + c_n^{(r)}x_n)}{\sum_{r=1}^N \left[ \prod_{i=1}^n \left( \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r} \right)^2 \right] \right) \right]}. \quad (47)$$

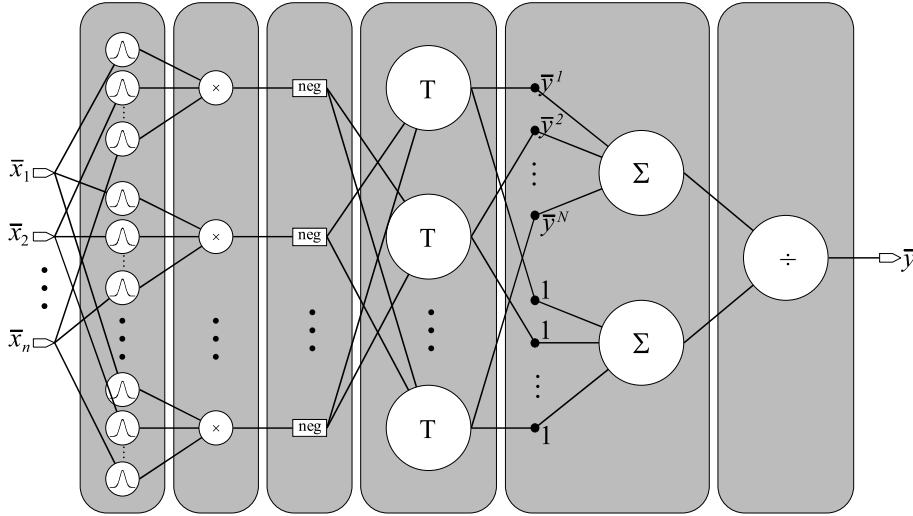
### Neuro-fuzzy Systems with Weights

We can distinguish fuzzy systems based on the following linguistic models:

- Systems based on linguistic model given by (14).
- Systems based on linguistic model given by (14), and additionally with weights  $w_k^{\text{agr}} \in [0, 1]$ ,  $k = 1, \dots, N$ , describing the importance of each rule:

$$R^{(k)}: \left[ \begin{array}{l} \text{IF } x_1 \text{ is } A_1^k \text{ AND } \dots \\ \text{AND } x_n \text{ is } A_n^k \\ \text{THEN } y \text{ is } B^k \end{array} \right] (w_k^{\text{agr}}). \quad (48)$$





Neuro-fuzzy Systems, Figure 4

Architecture of the logical neuro-fuzzy system with Lukasiewicz implication described by (41)

- (iii) Systems based on linguistic model given by (14), with weights  $w_k^{\text{agr}} \in [0, 1]$ ,  $k = 1, \dots, N$ , describing the importance of each rule and weights  $w_{i,k}^{\tau} \in [0, 1]$ ,  $k = 1, \dots, N$ ,  $i = 1, \dots, n$ , describing the importance of antecedents:

$$R^{(k)} : \left[ \begin{array}{l} \text{IF } x_1 \text{ is } A_1^k(w_{1,k}^{\tau}) \text{ AND } \dots \\ \text{AND } x_n \text{ is } A_n^k(w_{n,k}^{\tau}) \\ \text{THEN } y \text{ is } B^k \end{array} \right] (w_k^{\text{agr}}). \quad (49)$$

In order to incorporate the weights in models (48) and (49) into description of neuro-fuzzy systems, we can use the following formula (see [48,50]):

$$T^*\{a_1, \dots, a_n; w_1, \dots, w_n\} = \frac{n}{T} \{1 - w_i(1 - a_i)\}, \quad (50)$$

where  $T\{\cdot\}$  is any t-norm, and the weights meet the condition  $w_i \in [0, 1]$ ,  $i = 1, \dots, n$ . Observe that if we assume that  $w_i = 1$ ,  $i = 1, \dots, n$ , then function  $T^*\{\cdot\}$  is reduced to t-norm  $T\{\cdot\}$ . Analogously, we define

$$S^*\{a_1, \dots, a_n; w_1, \dots, w_n\} = \frac{n}{S} \{w_i a_i\}. \quad (51)$$

where  $S$  is any t co-norm.

**Example 7** Simplified neuro-fuzzy system (41) with Lukasiewicz fuzzy implication and weights describing the

importance of each rule takes the form

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \frac{n}{T^*} \left\{ 1 - \frac{n}{T} \left\{ \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right\}, w_k \right\}}{\sum_{r=1}^N \frac{n}{T^*} \left\{ 1 - \frac{n}{T} \left\{ \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right\}, w_k \right\}}. \quad (52)$$

**Example 8** Simplified neuro-fuzzy system (41) with Lukasiewicz fuzzy implication, weights describing the importance of each rule and weights describing the importance of antecedents takes the form

$$\bar{y} = \frac{\sum_{r=1}^N \bar{y}^r \frac{n}{T^*} \left\{ 1 - \frac{n}{T^*} \left\{ \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right], w_{i,k} \right\}, w_k \right\}}{\sum_{r=1}^N \frac{n}{T^*} \left\{ 1 - \frac{n}{T^*} \left\{ \exp \left[ - \left( \frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right], w_{i,k} \right\}, w_k \right\}}. \quad (53)$$

### Neuro-fuzzy Systems for Pattern Classification

We will explain how to modify linguistic model (13) to solve multi-classification problems. Let  $[x_1, \dots, x_n]$  be the vector of features of an object  $v$ . Let  $\Omega = \{\omega_1, \dots, \omega_M\}$  be a set of classes. The knowledge is represented by a set of

$N$  rules in the form

$$R^{(k)} : \begin{cases} \text{IF} & x_1 \text{ is } A_1^k \quad \text{AND} \\ & x_2 \text{ is } A_2^k \quad \text{AND} \dots \\ & x_n \text{ is } A_n^k \\ \text{THEN} & v \in \omega_1(z_1^k), \\ & v \in \omega_2(z_2^k), \dots \\ & v \in \omega_M(z_M^k), \end{cases}, \quad (54)$$

where  $z_j^k$ ,  $j = 1, \dots, M$ ,  $k = 1, \dots, N$ , are interpreted as “support” for class  $\omega_j$  given by rule  $R^{(k)}$ . We will now redefine description (54). Let us introduce vector  $\mathbf{z} = [z_1, \dots, z_M]$ , where  $z_j$ ,  $j = 1, \dots, M$ , is the “support” for class  $\omega_j$  given by all  $M$  rules. We can scale the support values to the interval  $[0,1]$ , so that  $z_j$  is the membership degree of an object  $v$  to class  $\omega_j$ , according to all  $M$  rules. The rules are represented by

$$R^{(k)} : \begin{cases} \text{IF} & x_1 \text{ is } A_1^k \quad \text{AND} \\ & x_2 \text{ is } A_2^k \quad \text{AND} \dots \\ & x_n \text{ is } A_n^k \\ \text{THEN} & z_1 \text{ is } B_1^k \quad \text{AND} \\ & z_2 \text{ is } B_2^k \quad \text{AND} \dots \\ & z_M \text{ is } B_M^k \end{cases}, \quad (55)$$

and formula (20) adopted for classification takes the form

$$\bar{z}_j = \frac{\sum_{k=1}^N \bar{z}_j^k \mu_{B_j^k}(\bar{z}_j^k)}{\sum_{k=1}^N \mu_{B_j^k}(\bar{z}_j^k)}, \quad (56)$$

where  $\bar{z}_j^k$  are centers of fuzzy sets  $B_j^k$ ,  $j = 1, \dots, M$ ,  $k = 1, \dots, N$ .

## Learning

Fuzzy and neuro-fuzzy system knowledge is expressed in the form of fuzzy rules (14). The learning process tries to pick the right number of rules and location of linguistic values (fuzzy sets defined in linguistic variables) to act like an expert or device performing given task. Traditional fuzzy systems are devised to work on the basis of knowledge formulated by an expert in the form of fuzzy rules in nearly natural language. The expert is not always able to formulate his or her knowledge or he or she is not available at all. In such cases we can rely on knowledge acquisition. The data acquired are used to supervised learning and then the learning process becomes similar to that known from

neural networks, especially gradient methods. To determine all parameters of neuro-fuzzy systems using numerical data, genetic and evolutionary algorithms can be also used. As knowledge in fuzzy systems is structured in the form of rules we can also employ other methods for learning from numerical data, e.g. clustering algorithms with most often used fuzzy c-means algorithm. We can also utilize any method for generating rules from data, e.g. [59].

Having given the learning dataset of the pair  $(\bar{\mathbf{x}}, d)$  where  $d$  is desired response of the system, we can use the following error measure:

$$Q(\bar{\mathbf{x}}, d) = \frac{1}{2} [\bar{y}(\bar{\mathbf{x}}) - d]^2. \quad (57)$$

Our goal is to minimize the error by determining the value of parameters  $\bar{x}_i^r$ ,  $\sigma_i^r$  (antecedent Gaussian fuzzy sets) and  $\bar{y}^r$  (consequent singleton). The parameters can be treated like weights in neural networks and in learning iteration (step)  $t$  the value of any weight can be defined by

$$p_i^k(t+1) = p_i^k(t) - \eta \frac{\partial Q(\bar{\mathbf{x}}, d; t)}{\partial p_i^k(t)}. \quad (58)$$

In the case of the simplest neuro-fuzzy system given by (37), we obtain the following formulas for gradients in recursive procedure (58)

$$\begin{aligned} \frac{\partial Q}{\partial \bar{x}_h^j} = & -2\varepsilon \frac{(\bar{x}_h^j - \bar{x}_h) \prod_{i=1}^n \mu_{A_i^j}(\bar{x}_i) \sum_{r=1}^N (\bar{y}^j - \bar{y}^r) \prod_{i=1}^n \mu_{A_i^r}(\bar{x}_i)}{\left\{ \sigma_h^j \sum_{r=1}^N \prod_{i=1}^n \mu_{A_i^r}(\bar{x}_i) \right\}^2}, \end{aligned} \quad (59)$$

$$\begin{aligned} \frac{\partial Q}{\partial \sigma_h^j} = & 2\varepsilon \frac{(\bar{x}_h^j - \bar{x}_h) \prod_{i=1}^n \mu_{A_i^j}(\bar{x}_i) \sum_{r=1}^N (\bar{y}^j - \bar{y}^r) \prod_{i=1}^n \mu_{A_i^r}(\bar{x}_i)}{\left\{ \sigma_h^j \sum_{r=1}^N \prod_{i=1}^n \mu_{A_i^r}(\bar{x}_i) \right\}^2}, \end{aligned} \quad (60)$$

$$\frac{\partial Q}{\partial \bar{y}^j} = \varepsilon \frac{\prod_{i=1}^n \mu_{A_i^j}(\bar{x}_i)}{\sum_{r=1}^N \prod_{i=1}^n \mu_{A_i^r}(\bar{x}_i)}, \quad (61)$$

where  $\varepsilon = \bar{y} - d$  is an error calculated in the last layer of the system and  $j = 1, \dots, n$ ,  $h = 1, \dots, N$ .

### Criteria Isolines Method

In this section, we will first present the Akaike criterion, initially applied to the order estimation of autoregression processes, and next it will be adapted to evaluate the effectiveness of neuro-fuzzy systems. By the effectiveness of operation of a neuro-fuzzy system, we shall understand the precision (accuracy) of operation achieved by such a system, (expressed by mean squared error or by the number of wrongly classified samples) in the context of its size. By the system size we shall understand the number of all parameters that are subject to learning. We shall also present the concept of the so-called criteria isolines, which allow one to solve the problem of the compromise between the system accuracy and the number of parameters describing this system.

Neuro-fuzzy system should be characterized by the smallest possible error but at the same time should be as simple as possible. It should be remembered that systems with smaller number of trained parameters are characterized among others by better generalization capabilities. Here, we should mention the so-called parsimony principle [52]. This principle is very useful when determining the appropriate order of the model. It may be formulated as follows: from between two alternative and satisfactory models, we shall choose the one which contains less independent parameters. This principle remains compliant with common sense: "Do not enter any additional parameters into the process description unless they are necessary".

Estimation methods of the system order have been best developed for autoregression processes [25,35]. Time series  $u(n), u(n-1), \dots, u(n-p)$  is an autoregression process of order  $p$ , if the difference equation is satisfied

$$u(n) + \alpha_1 u(n-1) + \dots + \alpha_p u(n-p) = e(n) \quad (62)$$

or equivalently

$$u(n) = - \sum_{k=1}^p \alpha_k u(n-k) + e(n), \quad (63)$$

where  $\alpha_1, \dots, \alpha_p$  are process coefficient, while  $e(n)$  is the white noise

$$E[e(n)e(m)] = \begin{cases} \sigma^2 & \text{dla } n = m \\ 0 & \text{dla } n \neq m \end{cases}. \quad (64)$$

In the autoregression theory, criteria allowing one to estimate the order of predictor  $p$ , determining first the prediction error  $\hat{Q}_p$  based on the learning sequence of the length  $M$ , are well known. The most important is the Akaike information criterion (AIC), Schwarz method and the final prediction error (FPE) method [25].

The Akaike estimation order prediction method will be adapted now to the evaluation of fuzzy systems. Thanks to this, search for the desired fuzzy system based on two criteria (number of parameters and mean square error) will come down to one selected criterion, i.e. AIC. It has been adapted for the needs of evaluation of neuro-fuzzy systems in the following form:

$$AIC(p, \hat{Q}_p) = M \ln \hat{Q}_p + 2p, \quad (65)$$

where  $p$  is the number of system parameters subject to learning (number of parameters of all membership functions and number of all weights if they occur in a given system),  $\hat{Q}_p$  is the measure of error used in simulations, and  $M$  is the number of samples in a learning sequence. The product  $M \cdot n$  may, therefore, be treated as a measure of size of the problem being solved. Table 2 contain the computed values of criteria for particular tested structures in case of the learning and testing sequence used in the polymerization problem [56]. Figure 5 illustrates the coordinates of the points corresponding to particular neuro-fuzzy systems tested. The coordinate  $p$  defines the number of parameters of a given system; coordinate  $Q$  defines the error with which the system realized the problem to be solved.

The criteria isolines present constant values of the AIC criterion, with different values of the error and the number of parameters. Such an approach allows one to solve the problem of the compromise between the system operation error and the number of parameters describing this system. Points located on the criteria isolines with the same values of AIC criterion characterize the neuro-fuzzy systems making up the Pareto set. In the Pareto set, none of the two values of contradictory criteria may be improved (mean square error versus system size), without worsening the other one. Points located on the criteria isolines with the smallest values of AIC criterion characterize the neuro-fuzzy systems which have been called suboptimal ones. The suboptimal neuro-fuzzy systems presented in graphs ensure the smallest value of criteria within tested structures (the terminology "optimum systems" is not used as all possible structures have not been tested). From Table 2 it follows that for both learning and testing sequences the best, in the sense of the Akaike criterion, is the structure no. 1 (Larsen simplified) followed by the structure no. 29 (Zadeh with weights of rules).

### Future Directions

Mankind creates and collects large volumes of data nowadays. We deal with an enormous amount of, often multidimensional, data. We can find such data in medicine

Neuro-fuzzy Systems, Table 2

Value of the Akaike criterion for the polymerization problem for the learning and testing sequence

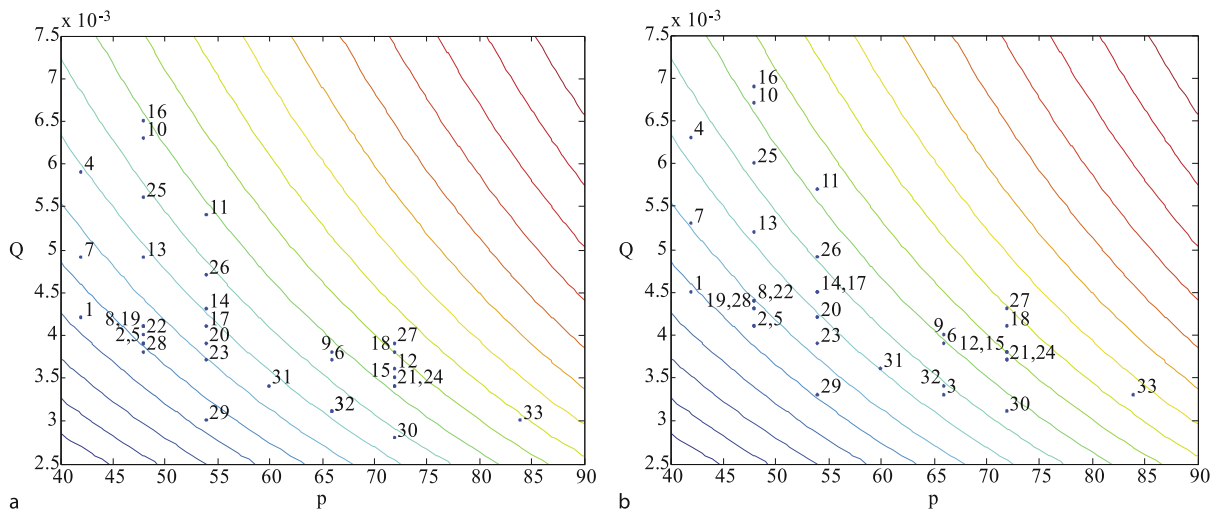
No.	Structure	Polymerization				
		p	Learning sequence		Testing sequence	
			Error	AIC	Error	AIC
1	Larsen simplified	42	0.0042	−299.09	0.0045	−294.26
2	Larsen simplified with weights of rules	48	0.0039	−292.27	0.0041	−288.77
3	Larsen simplified with weights of inputs and rules	66	0.0031	−272.34	0.0033	−267.97
4	Łukasiewicz simplified	42	0.0059	−275.30	0.0063	−270.70
5	Łukasiewicz simplified with weights of rules	48	0.0039	−292.27	0.0041	−288.77
6	Łukasiewicz simplified with weights of inputs and rules	66	0.0037	−259.96	0.0039	−256.27
7	Zadeh simplified	42	0.0049	−288.30	0.0053	−282.80
8	Zadeh simplified with weights of rules	48	0.0041	−288.77	0.0044	−283.83
9	Zadeh simplified with weights of inputs and rules	66	0.0038	−258.09	0.0040	−254.50
10	Binary	48	0.0063	−258.70	0.0067	−254.40
11	Binary with weights of rules	54	0.0054	−257.49	0.0057	−253.71
12	Binary with weights of inputs and rules	72	0.0036	−249.88	0.0038	−246.09
13	Larsen	48	0.0049	−276.30	0.0052	−272.14
14	Larsen with weights of rules	54	0.0043	−273.44	0.0045	−270.26
15	Larsen with weights of inputs and rules	72	0.0035	−251.85	0.0038	−246.09
16	Łukasiewicz	48	0.0065	−256.52	0.0069	−252.34
17	Łukasiewicz with weights of rules	54	0.0041	−276.77	0.0045	−270.26
18	Łukasiewicz with weights of inputs and rules	72	0.0038	−246.09	0.0041	−240.77
19	Mamdani	48	0.0041	−288.77	0.0043	−285.44
20	Mamdani with weights of rules	54	0.0039	−280.27	0.0042	−275.09
21	Mamdani with weights of inputs and rules	72	0.0034	−253.88	0.0037	−247.96
22	Reichenbach	48	0.0040	−290.50	0.0044	−283.83
23	Reichenbach with weights of rules	54	0.0037	−283.96	0.0039	−280.27
24	Reichenbach with weights of inputs and rules	72	0.0034	−253.88	0.0037	−247.96
25	Willmott	48	0.0056	−266.95	0.0060	−262.12
26	Willmott with weights of rules	54	0.0047	−267.21	0.0049	−264.30
27	Willmott with weights of inputs and rules	72	0.0039	−244.27	0.0043	−237.44
28	Zadeh	48	0.0038	−294.09	0.0043	−285.44
29	Zadeh with weights of rules	54	0.0030	−298.64	0.0033	−291.97
30	Zadeh with weights of inputs and rules	72	0.0028	−267.47	0.0031	−260.34
31	Takagi–Sugeno	60	0.0034	−277.88	0.0036	−273.88
32	Takagi–Sugeno with weights of rules	66	0.0031	−272.34	0.0034	−265.88
33	Takagi–Sugeno with weights of inputs and rules	84	0.0030	−238.64	0.0033	−231.97

and health care, business, manufacturing and finance, science and telecommunication. Neuro-fuzzy systems have to cope with these large, often multidimensional and incomplete, data sets.

In case of real world data, learning systems have to face with a problem of missing data, i.e. missing certain features (input variables). Such shortages can be replaced by the average or modal value, by a value provided by the user or by a value resulting from other records. Good solution in this case is provided by information systems based on rough sets. Such systems work very similarly to classification made by humans – having data with missing values

they perform classification, increasing at the same time decision uncertainty level. In case of lack of even one feature, fuzzy systems do not work properly. Fusion of fuzzy and rough sets seems to be a promising field and in the near future rough-neuro-fuzzy systems are expected to emerge.

The development of single learning systems comes slowly to a saturation point and one of ways for further development is creating ensembles of these systems. The accuracy and generalization ability is nearly always improved comparing with a single system. Learning systems can be combined using different learning data, different systems and different aggregation methods. There are sev-



**Neuro-fuzzy Systems, Figure 5**

**Criteria isolines: results obtained by particular systems for the Akaike criterion for the polymerization problem: a learning sequence, b testing sequence**

eral methods for creating ensembles of learning systems. The most popular are bagging and boosting which are meta-algorithms allowing use of nearly any learning for ensemble members. After adding a new member to the ensemble, new learning data are created taking into account learning efficiency for every learning sample. In consecutive steps samples which are difficult to learn have more influence on learning. In the most popular AdaBoost algorithm the influence is achieved by assigning weights to learning samples according to their previous performance. As a result of learning of such an ensemble, one will obtain several fuzzy rule bases. To merge the rule bases into one knowledge base special algorithms for simplification of such large fuzzy rule bases should be developed.

The engineering community has accepted fuzzy logic and neuro-fuzzy systems for use in home appliances, automotive industry, robotics, cameras, air conditioning, image recognition, financial engineering, etc. The number and diversity of the applications is still growing and still there are many challenging problems to be solved.

### Acknowledgments

This work was supported in part by the Foundation for Polish Science (Professorial Grant 2005–2008) and the Polish Ministry of Science and Higher Education (Special Research Project 2006–2009 and Polish-Singapore Research Project 2008–2010) and by science funds for 2007–2010 as research project No. N N516 1669 33 and No. N N516 1155 33.

### Bibliography

1. Aliev RA, Aliev RR (2001) Soft computing and its applications. World Scientific Publishing, Singapore
2. Berenji HR, Khedkar P (1992) Learning and tuning fuzzy logic controllers through reinforcements. *IEEE Trans. Neural Networks*, vol 3, pp 724–740, October
3. Bubnicki Z (2001) Uncertain variables and their application to decision making. *IEEE Trans. on SMC, Part A: Systems and Humans*, vol 31, pp 587–596
4. Bubnicki Z (2002) Uncertain logics, variables and systems. Springer, Berlin-London-New York
5. Bubnicki Z (2002) A unified approach to descriptive and prescriptive concepts in uncertain decision systems, *Systems Analysis Modeling Simulation*, vol 42, issue 3. Gordon and Breach Science Publishers, Newark, pp 331–342
6. Chen MY, Linkens DA (2001) A systematic neuro-fuzzy modeling framework with application to material property prediction. *IEEE Trans. on Fuzzy Systems*, vol 9, pp 781–790
7. Chuang CC, Su SF, Chen SS (2001) Robust TSK fuzzy modeling for function approximation with outliers. *IEEE Trans. on Fuzzy Systems*, vol 9, pp 810–821, December
8. Corcoran AL, Sen S (1994) Using real-valued genetic algorithms to evolve rule sets for classification. *Proc. of the 1st IEEE Conf. Evolut. Computat.*, Orlando FL, June pp 120–124
9. Czogała E, Łęski (2000) J fuzzy and neuro-fuzzy intelligent systems, Physica-Verlag Company, Heidelberg, New York
10. Duan JC, Chung FL (2001) Cascaded fuzzy neural network based on syllogistic fuzzy reasoning. *IEEE Trans. on Fuzzy Systems*, vol 9, pp 293–306
11. Eubank RL (1999) Nonparametric regression and spline smoothing. Marcel Dekker, New York
12. Fodor JC (1991) On fuzzy implication operators, *Fuzzy Sets and Systems*, vol 42, issue 3. Elsevier, Amsterdam, pp 293–300
13. Fogel DB (1995) Evolutionary computation: towards a new philosophy of machine intelligence. IEEE Press, New York
14. Fuller R (2000) Introduction to neuro-fuzzy systems, advances in soft computing. Physica-Verlag, New York



15. Gaweda AE, Żurada JM (2000) Fuzzy neural network with relational fuzzy rules. *Proc. of the Intern. Joint Conference on Neural Networks IJCNN'2000*, vol 5, pp 3–8, Como, Italy, July 23–27
16. Gaweda AE, Żurada JM (2001) Data-driven design of fuzzy system with relational input partition. *Proc of the Int Conference on Fuzzy Systems FUZZ-IEEE'2001*, Melbourne, Australia, December 2–5
17. Hirota K (1993) *Industrial Applications of Fuzzy Technology*. Springer, Tokyo, Berlin, Heidelberg, New York
18. Jang JSR (1993) ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. Syst., Man, Cybern.*, vol 23, pp 665–685, June
19. Jang JSR, Sun CT (1995) Neuro-fuzzy modeling and control. *Proc IEEE*, vol 83, pp 378–406, March
20. Jang JS, Sun CT, Mizutani E (1997) *Neuro-fuzzy and soft computing*. Prentice Hall, Englewood Cliffs
21. Juang C-F, Lin C-T (1998) An on-line self-constructing neural fuzzy inference network and its applications. *IEEE Trans. on Fuzzy Systems*, vol 6, pp 12–32, February
22. Kacprzyk J (1997) *Multistage fuzzy control*. Wiley, Chichester
23. Kasabov N (1996) *Foundations of neural networks, fuzzy systems and knowledge engineering*. The MIT Press CA, Cambridge
24. Kasabov N (2002) DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction. *IEEE Trans. on Fuzzy Systems*, vol 10, pp 144–154, April
25. Kay SM (1988) *Modern spectral estimation. Theory and application*. Prentice Hall, Englewood Cliffs
26. Kecman V (2001) *Learning and soft computing*. MIT Press, Cambridge
27. Klement EP, Mesiar R, Pap E (2000) *Triangular norms*. Kluwer, Dordrecht
28. Lee K-M, Kwak D-H, Lee-Kwang H (1994) A fuzzy neural network model for fuzzy inference and rule tuning. *Int J Uncertainty, Fuzziness and Knowledge-Based Systems*, vol 2, no. 3, pp 265–277
29. Lee K-M, Kwak D-H, Lee-Kwang H (1996) Fuzzy inference neural network for fuzzy model tuning. *IEEE Trans on Systems, Man, and Cybernetics. Part B*, vol 26, No. 4, pp 637–645
30. Lin CT (1994) *Neural fuzzy control systems with structure and parameter learning*. World Scientific, Singapore
31. Lin CT, Lee CSG (1991) Neural-network-based fuzzy logic control and decision system. *IEEE Trans Comput*, vol 40, pp 1320–1336, December
32. Lin CT, Lee GCS (1997) *Neural fuzzy systems a neuro-fuzzy synergism to intelligent systems*. Prentice Hall, Englewood Cliffs
33. Lin CT, Lu YC (1995) A neural fuzzy systems with linguistic teaching signals. *IEEE Trans on Fuzzy Systems*, vol 3, pp 169–189, May
34. Lin Y, Cunningham GA III (1995) A new approach to fuzzy-neural system modeling. *IEEE Trans. on Fuzzy Systems*, vol 3, pp 190–198, May
35. Marple SL Jr (1987) *Digital spectral analysis with applications*. Prentice Hall, Englewood Cliffs
36. Michalewicz Z (1992) *Genetic algorithms + data structures = evolution programs*. Springer, Berlin
37. Mouzouris GC, Mendel JM (1997) Nonsingleton fuzzy logic systems: Theory and application. *IEEE Trans on Fuzzy Systems*, vol 5, No. 1, pp 56–71
38. Nauck D, Kruse R (1996) Designing neuro-fuzzy systems through back-propagation. In: Pedrycz W (ed) *Fuzzy modeling: paradigms and practice*. Kluwer, Boston, pp 203–228
39. Nauck D, Kruse R (1999) *Neuro-fuzzy systems for function approximation*. Fuzzy Sets and Systems, vol 101, pp 261–271
40. Nauck D, Klawon F, Kruse R (1997) *Foundations of neuro-fuzzy systems*. Wiley, Chichester
41. Nie J, Linkens D (1995) *Fuzzy-Neural Control. Principles, algorithms and applications*. Prentice Hall, New York, London
42. Pagan A, Ullah A (1999) *Nonparametric econometrics*. Cambridge Univ. Press, London
43. Pawlak Z (1982) Rough sets. *Int J Inform Comput Sci*, vol 11, no. 341
44. Pawlak Z (1991) *Rough sets. Theoretical aspects of reasoning about data*. Kluwer, Dordrecht
45. Pedrycz W (1992) Fuzzy neural networks with reference neurons as pattern classifiers. *IEEE Trans Neural Networks*, vol 3, no. 5, pp 770–775
46. Roubos H, Setnes M (2001) Compact and transparent fuzzy models and classifiers through iterative complexity reduction. *IEEE Trans. on Fuzzy Systems*, vol 9, pp 516–524, August
47. Rutkowska D (2002) *Neuro-fuzzy architectures and hybrid learning*. Springer, Heidelberg
48. Rutkowski L (2004) *Flexible neuro-fuzzy systems*. Kluwer, Norwell
49. Rutkowski L, Cpałka K (2000) Flexible structures of neuro-fuzzy systems, *Quo Vadis Computational Intelligence, Studies in Fuzziness and Soft Computing*, vol 54. Springer, Berlin, pp 479–484
50. Rutkowski L, Cpałka K (2003) Flexible neuro-fuzzy systems. *IEEE Trans. Neural Networks*, vol 14, pp 554–574
51. Rutkowski L, Rafajłowicz E (1989) On global rate of convergence of some nonparametric identification procedures. *IEEE Trans. on Automatic Control*, vol AC-34, no.10, pp 1089–1091
52. Söderström T, Stoica P (1989) *System identification*. Prentice-Hall, London
53. Tadeusiewicz R (1993) *Neural networks RM*. Academic Publishing House, Warsaw (in Polish)
54. Tadeusiewicz R (1998) *Elementary introduction to neural networks with computer programs*. Academic Publishing House, Warsaw (in Polish)
55. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. *IEEE Trans Systems, Man, and Cybernetics*, vol 15, pp 116–132
56. UCI respository of machine learning databases, Available online: <http://ftp.ics.uci.edu/pub/machine-learning-databases/>
57. Wang JS, Lee CSG (2002) Self-adaptive neuro-fuzzy inference systems for classification applications. *IEEE Trans. on Fuzzy Systems*, vol 10, pp 790–802
58. Wang LX (1994) *Adaptive Fuzzy Systems and Control*. PTR Prentice Hall, Englewood Cliffs
59. Wang LX, Mendel JM (1992) Generating fuzzy rules by learning from examples, *IEEE Transactions on Systems. Man and Cybernetics*, vol 22, no. 6, pp 1414–1427
60. Yager RR (1990) Fuzzy logic controller structures. *Proc. SPIE Symp. Laser Sci. Optics Appl.* 368–378
61. Yager RR (1992) A general approach to rule aggregation in fuzzy logic control. *Appl Intelligence*, vol 2, pp 333–351
62. Zadeh LA (1965) Fuzzy sets. *Information and Control*, vol 8, no. 3, pp 338–353
63. Żurada JM (1992) *Introduction to artificial neural systems*. West Publishing Company

## Neuronal Dynamics

NICOLAS BRUNEL<sup>1,2</sup>, VINCENT HAKIM<sup>3</sup>

<sup>1</sup> Laboratoire de Neurophysique et Physiologie,  
Université Paris Descartes, Paris, France

<sup>2</sup> UMR 8119, CNRS, Paris, France

<sup>3</sup> Laboratoire de Physique Statistique, UMR 8550, CNRS,  
Paris, France

### Article Outline

Glossary

Definition of the Subject

Introduction

Types of Modeling and Theoretical Tools

Intrinsic Network Dynamics

Stimulus Driven Dynamics

Future Directions

Bibliography

### Glossary

**Action potential** or Spikes: electrical pulses of an amplitude of about 100 mV that travel along nerve fibers.

**Field potential** an electrical signal recorded extracellularly which arises from the synchronized activity of many cells.

**Hippocampus** one of the most studied area of the mammalian nervous system which is part of the limbic system and is involved in learning and memory.

**Neuron** the main excitable cells of nerve tissue.

**Network** an ensemble of synaptically connected cells.

**Synapse** the specialized junction between two neurons where the action potential voltage transient in the presynaptic cell is transmitted to the post-synaptic cell via neurotransmitter release (chemical synapse) or direct electrical connection (electrical synapse).

### Definition of the Subject

How information is processed by nervous systems is a question of major interest, with far-reaching implications for domains as diverse as medicine and philosophy. It is however still far from understood despite much work and progress during the last sixty years. Neurons are cells which exhibit diverse dynamical behaviors. Our present partial view makes it clear that, besides physiology of single neurons and anatomy of neural systems, understanding the dynamics of coupled neuron assemblies and their collective dynamics is of utmost importance. This requires supplementing the tools of experimental biology,

that themselves are making impressive progress, by modeling and theoretical analysis.

### Introduction

Since the first electrical recordings, it has been noted that the mammalian brain exhibits diverse patterns of activity [14]. Oscillations at various frequencies are prominent features of human electroencephalograms that is, voltage signals recorded from the scalp [19]. Their dominant frequencies are state and task dependent shifting from slow “delta” frequencies (1–4 Hz) in certain stages of sleep, to “alpha” frequencies (8–13 Hz) in quiet wakefulness or “beta” frequencies (13–30 Hz) in attentive immobility. Intracranial recordings of field potential or neuron activity in animals have revealed higher frequency components from 40 Hz to more than 200 Hz and have also shown that neural rhythms depend on the neural structures from which they originate [3]. For instance, in the rat hippocampus, rhythms at theta (4–8 Hz) and gamma (30–100 Hz) frequencies are observed during exploratory behaviors whereas intermittent fast oscillations (100–200 Hz) are recorded during awake immobility and consummatory behaviors [34]. In the rat olfactory bulb, odorants induce strong gamma oscillations [4].

Electrophysiological recordings, visualization, pharmacological and genetic manipulations are important experimental tools for observing and assessing the mechanisms underlying this complex neural activity. However, analyzing the dynamics of a large ensemble of coupled dynamical elements is a difficult task. Modeling, theoretical analysis and computer simulations are proving increasingly helpful to interpret experimental data, as it becomes more precise and extensive, and to direct further experiments.

In the following, we first review the different types of models that are currently most useful in analyzing network dynamics. We then consider various topics to which theoretical modeling has significantly contributed. We start in Sect. “[Intrinsic Network Dynamics](#)” by discussing spontaneous neural activity, that is, activity not directly induced by stimulus-processing. Experimentally, “background” spontaneous activity appears to consist of neurons firing irregularly and at low rates and poses a first modeling challenge. The coexistence of different attractors, a question of central interest in different contexts, is considered next. We then review the different mechanisms that are thought to give rise to the various time scales of neural dynamics. In Sect. “[Stimulus Driven Dynamics](#)”, we examine how stimulus presentation can modify neural dynamics and discuss various ways in which

dynamics can lead to information processing. We begin by discussing two contrasting views: classically, sensory processing has been thought to proceed by a feedforward and stimulus-driven dynamic, whereas some more recent modeling and experiments suggest that stimuli provide only a weak bias on endogenous dynamics. We then examine various other ways in which dynamics can contribute to information processing and memory. The precise role of oscillations at diverse frequencies in information processing is still ill-understood. We briefly discuss two of the better understood structures and several proposals. We finally conclude, in Sect. “Future Directions”, by pointing out some approaches that appear to bear promises of future progress.

### Types of Modeling and Theoretical Tools

The first task of a modeler in computational or theoretical neuroscience is to specify a particular model to be investigated. Choosing which model depends, of course, on the nature of the problem one is interested in, as well as on the availability of relevant experimental data. For example, if one is interested in understanding the origin of the irregularity of spike trains in cortex as observed in extracellular recordings in vivo, it is clear that one needs a spiking neuron model. For people working at the network level, this means specifying at least three things: the ‘architecture’ of the network (number of ‘units’, a specification of who communicates with who, and finally specification of the inputs); a model for the ‘units’ themselves (these ‘units’ can be groups or populations of functionally equivalent neurons, or single neurons); and a model for the connections between the units (synapses in the case units are single neurons). In this section, we first briefly describe popular models of units and connections, and then commonly studied network architectures. Finally, we explain briefly the theoretical methods that are used to investigate such networks.

### Different Classes of Neurons Models

Broadly speaking, there are three classes of models: rate models; networks of binary neurons; and networks of spiking neurons. Though networks of binary neurons are a very useful tool to investigate associative memory properties of recurrent networks, they are not well-suited to study questions involving dynamics. We therefore restrict ourselves here to discussing rate models and networks of spiking neurons.

**Rate Models** In so-called ‘rate models’ (also called neural mass models, or neural field models), one describes the

activity of a population of functionally equivalent neurons by a continuous variable  $r$  that typically obeys an ordinary differential equation. In its simplest form, the activity of a single population of neurons can be described by the equation

$$\tau \frac{dr}{dt} = -r + \Phi(I_{\text{ext}} + Jr), \quad (1)$$

where:  $r$  is the mean activity (firing rate) of the population,  $\tau$  is the characteristic time constant of rate dynamics,  $\Phi$  is the steady state input-output transfer function (f-I curve), that is typically taken to be sigmoidal,  $I_{\text{ext}}$  are the external inputs to the population, and  $J$  is the strength of the intra-population connections. For excitatory populations,  $J > 0$ , while for inhibitory populations,  $J < 0$ .

The function  $\Phi$  that describes the ‘static’ transfer function of neurons in the population can take different forms. Popular transfer functions are threshold-linear,

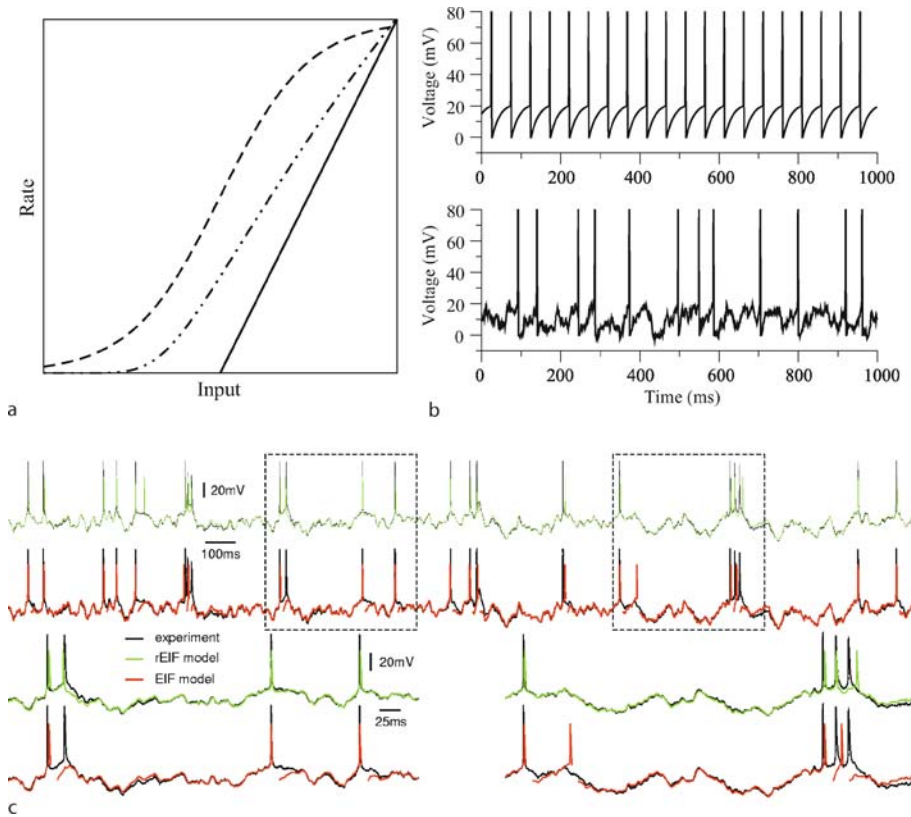
$$\Phi(I) = \begin{cases} I - T & I > T \\ 0 & I < T \end{cases}, \quad (2)$$

where  $T$  is a threshold, or sigmoidal,  $\Phi(I) = r_{\text{max}}/(1 + \exp(-\beta(I - T)))$ , where  $T$  is again a threshold and  $\beta$  measures the steepness of the f-I curve at half-maximal firing rate. Note that while the threshold-linear transfer function is unbounded, the sigmoidal one saturates at a maximal rate  $r_{\text{max}}$ . In the limit  $\beta \rightarrow \infty$ , the sigmoidal transfer function becomes binary,  $\Phi = r_{\text{max}}$  for  $I > T$ ,  $\Phi = 0$  otherwise.

Another choice is to take  $\Phi$  as the analytically computed transfer function of a specific spiking neuron model, like the integrate-and-fire model. As described below, one can compute the average firing frequency as a function of both mean and variance of inputs in some simple cases, and the resulting function can be used as a transfer function in a rate model formulation [7]. Examples of f-I curves are shown schematically in Fig. 1a.

**Spiking Neurons** As for most cells, ionic concentration differences and selective membrane permeability make neurons intracellularly hyperpolarized with respect to the extracellular medium. The difference between the intracellular and extracellular potentials, the cell membrane potential, is thus negative, with a typical value of about  $-70$  mV. Most neurons in the nervous systems of vertebrates use action potentials<sup>1</sup>, or spikes, which are large depolarizations of short duration ( $\sim 1$  ms), as communication units. It is therefore natural to use spiking neuron

<sup>1</sup>Note however that there are well-documented exceptions, for instance in the retina.



Neuronal Dynamics, Figure 1

Single unit/neuron models. **a** Rate model: commonly used  $f$ - $I$  curves  $\Phi$  (mean firing rate as a function of inputs). *Solid*: threshold-linear function. *Dashed*: sigmoidal function. *Dot-dashed*:  $f$ - $I$  curve of a leaky integrate-and-fire neuron subjected to noisy inputs. **b** Dynamics of a leaky integrate-and-fire neuron (membrane potential) driven by constant suprathreshold input (*upper panel*) or noisy input (*lower panel*). Spikes are added by hand to better visualize the timing. **c** Dynamics of two versions of an exponential integrate-and-fire model (*red*: with hard reset, *green*: with adaptive threshold), together with recording of a cortical pyramidal cell. Both the artificial and the real cell receive the same injected time-dependent current (a realization of an Ornstein-Uhlenbeck process). From [8]

models in network studies. The action potential is primarily a membrane phenomenon that can be described with different degrees of realism and refinement. There is a correspondingly large diversity of single neuron spiking models in the literature, from simple, one-variable integrate-and-fire models, to complex multi-compartmental models with many voltage-gated ionic channels. The complexity of a single neuron model can be characterized along two dimensions: (i) the number of variables describing the membrane dynamics at a given location; and (ii) the number of compartments describing the spatial geometry of the neuron.

Along the first dimension, the simplest model is the integrate-and-fire model. It retains as its only variable the membrane potential [27,28,68,74,112] and simply describes the action potential by a threshold  $V_T$  in membrane potential. In the subthreshold range  $V < V_T$ , the

membrane dynamic is simply described as that of a capacitance  $C$  (of about  $1 \mu\text{F}/\text{cm}^2$ , coming from the insulating property of the lipid bilayer) in parallel with a passive leak conductance  $g$ ,

$$\tau_m \frac{dV}{dt} = -(V - V_L) + I_{\text{syn}}(t), \quad (3)$$

where  $\tau_m = C/g$  is the membrane time constant (typical values are 10–20 ms),  $V_L$  is the resting potential, and  $I_{\text{syn}}(t)$  are the synaptic inputs, in which we have absorbed the input resistance of the neuron (typical values are 10–100  $\text{M}\Omega$ ). Thus, in Eq. (3), the inputs are in mV. In this model, a spike is emitted when the voltage hits the threshold  $V_T$ . At such times, the voltage is immediately reset to a sub-threshold voltage  $V_R$ . It is also possible to add to the model an absolute refractory period, during which the voltage is clamped to the reset, to get saturation of the



firing frequency. To choose the values of the voltage parameters, there are basically two options: take the ‘realistic’ values (similar to observed values in real neurons),  $V_L \sim -70$  mV,  $V_T \sim -50$  mV, and  $V_R$  somewhere in between; or define the origin of the voltage to be  $V_L = 0$ , and redefine  $V_T$  and  $V_R$  accordingly. The dynamic of the integrate-and-fire model to various types of inputs is shown in Fig. 1b.

At the opposite extreme, ‘biophysically realistic’ models aim to accurately represent the dynamics of the ionic channels that modify membrane permeability. They often employ the celebrated Hodgkin–Huxley formalism [55] which describes the ionic channel contribution as changes in membrane conductance via the dynamics of the activation and inactivation of various voltage-gated ionic currents. While the integrate-and-fire model, due to its very simplicity, is a model of choice for analytical studies and large-scale simulations, it lacks many features exhibited by real neurons. These features can often be accurately captured in the Hodgkin–Huxley formalism but at the price of difficulties in mathematical analysis (due to a large number of variables and pronounced non-linearities) and heavy computational cost. An intermediate class of models try to capture the best of both worlds: they contain the minimal number of variables (typically 1 or 2) and the minimal non-linearity required to capture particular phenomena of interest [21]. For example, generalized integrate-and-fire neurons contain an additional variable coupled to the voltage, which can give rise to sub-threshold resonance, as seen in many cell types [61, 104]; non-linear integrate-and-fire models can accurately capture spike generation with a minimal (exponential) non-linearity in the voltage equation [8, 42], see Fig. 1c; firing rate adaptation can be captured by a single adaptation variable coupled to the voltage; and so on.

Along the second (spatial) dimension, a similar diversity exists. The simplest models are composed of a single compartment. At the opposite, detailed models of cells with highly branched dendritic trees, such as the cerebellar Purkinje cell, can contain more than a thousand compartments [36]. Again, intermediate models, such as two-compartmental models, try to capture the best of both worlds: capturing non-trivial phenomena exhibited by real cells with the minimal description.

While multi-compartment models have been a valuable tool at the single neuron level, most studies of network behaviors use neurons with a single compartment. Apart from the obvious issues of mathematical tractability and computational cost, another concern limiting the use of such models is the limited present experimental knowledge on relative geometry and repartition of ionic

channels. Ongoing initiatives, such as the “blue brain” project [86], may lead this to change, but we expect progress in this direction to be slow, given the huge number of parameters (especially if there are correlations in the properties of connected neurons).

## Synapses

While in rate models synapses are specified by a single number, in spiking neuron models synapses are specified by dynamical variables. In the simplest case, synapses have static amplitude: their response to a particular pre-synaptic spike is the same regardless of the history of their activation. However, many, if not all, synapses in the CNS are history-dependent. This is described by models capturing short- and/or long-term plasticity properties.

**Total Synaptic Inputs** A typical neuron in the CNS receives inputs from something on the order of 10,000 presynaptic neurons<sup>2</sup>. The total synaptic inputs  $I_{\text{syn}}(t)$  of Eq. (3) are typically taken as a linear sum of both individual synaptic inputs and individual spikes. This leads to a synaptic current to neuron  $i$  of the type

$$I_i(t) = \sum_j J_{ij} \sum_k s(t - t_j^k),$$

where the sum over  $j$  runs over pre-synaptic neurons,  $J_{ij}$  defines the amplitude of a unitary post-synaptic potential (PSP) elicited by a spike of neuron  $j$  in neuron  $i$  ( $J_{ij} > 0$  if neuron  $j$  is excitatory, while  $J_{ij} < 0$  if neuron  $j$  is inhibitory), the sum over  $k$  is a sum over spikes of pre-synaptic neuron  $j$ ,  $s$  describes the temporal dynamics of a unitary post-synaptic current (PSC see below), and  $t_j^k$  is the timing of the  $k$ th spike emitted by neuron  $j$ .

The simplest models use the so-called “current-based” description of synaptic inputs, in which the current is independent of the voltage. A more realistic description is to take a voltage dependent  $J \propto g(V - V_{\text{rev}})$  where  $g$  is the time-dependent synaptic conductance, and  $V_{\text{rev}}$  is the corresponding reversal potential – the “conductance-based” description. Typical values for PSP amplitudes are in the range 0.1–1 mV;  $V_{\text{rev}} \sim 0$  mV for excitatory synapses, while  $V_{\text{rev}} \sim -80$  mV for inhibitory synapses. Typical values for synaptic conductances are in the nS range.

## A Simplified Semi-realistic Description for Single PSCs

The time course of post-synaptic currents elicited by a sin-

<sup>2</sup>This is an estimate for pyramidal cells in neocortex and hippocampus. Different cell types can have widely different average number of inputs, from the 4 average inputs of a granule cell in cerebellum, to the more than 100,000 inputs of a Purkinje cell.



gle spike can be described quite accurately by a delayed difference of exponentials. When a spike is emitted at time  $t = 0$ , the current is zero until  $t = D$ , where  $D$  is the delay, or latency, of the synaptic connection. Then the current is described by

$$s(t) \propto \exp\left(-\frac{(t-D)}{\tau_d}\right) - \exp\left(-\frac{(t-D)}{\tau_r}\right), \quad (4)$$

where  $\tau_r < \tau_d$  describe the rise and decay times of the current, respectively. This can be equivalently described by the system of differential equations

$$\begin{aligned} \tau_d \frac{ds}{dt} &= -s + x \\ \tau_r \frac{dx}{dt} &= -x + \delta(t - D). \end{aligned}$$

Typical values for the time constants are:  $D \sim 1$  ms;  $\tau_r$  and  $\tau_d$  vary widely depending on receptor type, but are  $\tau_r < 1$  ms,  $\tau_d \sim 2$ –10 ms for the fast (AMPA and GABA<sub>A</sub>) receptor-mediated synaptic responses. See Fig. 2 for synaptic currents described by Eq. (4), as well as a few experimentally recorded synaptic currents.

**Dynamic Synapses** Virtually all synapses in the CNS undergo plasticity at various time scales. Short-term plasticity (on timescales of 100s of ms) is ubiquitous; Short-term depression and facilitation has been characterized experimentally and models exist that reproduce fairly accurately the characteristics of this plasticity (see [2,117]).

On longer timescales ( $> 1$  hour) various types of synapses (for example, pyramidal-to-pyramidal synapses

in the neocortex; CA3 to CA1 synapses in the hippocampus; granule cell to Purkinje cell synapses in the cerebellum) exhibit various types of long-term plastic changes. There is a vast amount of material documenting such plasticity in the literature (for reviews see [16,82,83,119]), and a large number of investigators have proposed models that account for rate-based and/or spike-timing based plasticity.

In this review, we will mostly consider networks with fixed synapses; the effects of both short-term and long-term plasticity will be mentioned only briefly.

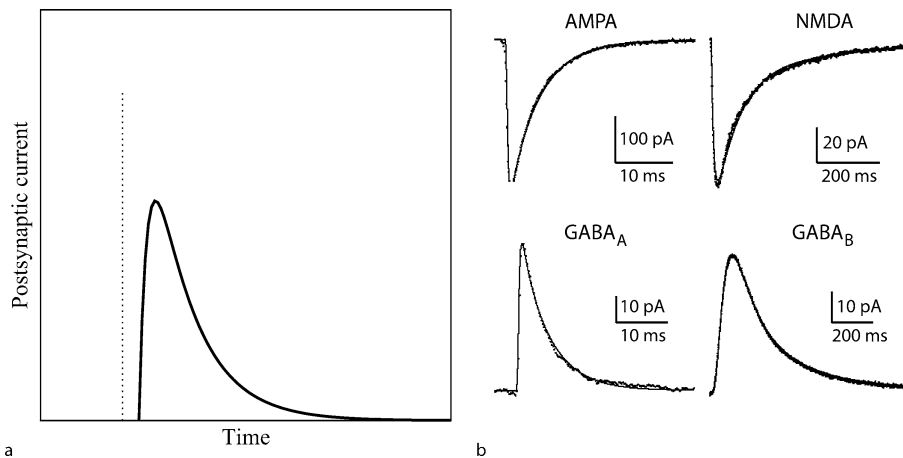
### Network Architecture

The network architecture specifies how the units, or neurons, are coupled together.

**Architectures in Rate Models** Rate models are often used to investigate the dynamics of spatially extended networks, or systems which encode continuous stimuli. A prototypical example is given by models of the primary visual cortex, in which the activity  $r(x)$  at a given location  $x$  evolves according to

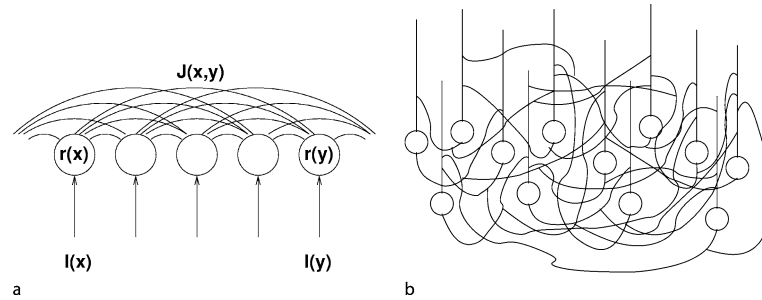
$$\tau \frac{dr(x)}{dt} = -r(x) + \Phi \left( I_{\text{ext}}(x, t) + \int dy J(x, y) r(y) \right). \quad (5)$$

The network ‘architecture’ is characterized by the ‘synaptic footprint’  $J(x, y)$  that specifies the strength of connections from  $y$  to  $x$  (the recurrent part); and the inputs  $I_{\text{ext}}(x, t)$ , describing inputs coming from outside the modeled net-



Neuronal Dynamics, Figure 2

**Synaptic models.** **a** Synaptic current elicited by a single spike of Eq. (4). Timing of the presynaptic spike is indicated schematically by a dotted line. **b** Synaptic currents mediated by various receptors, recorded in various slice preparations. From [38]



Neuronal Dynamics, Figure 3

**Network architectures.** **a** Sketch of a rate model. Each circle represents a population of cells at a given location. Lines connecting circles represent synaptic connections between populations. **b** A randomly connected network of individual neurons. The drawing schematically represents cell bodies (*circles*), dendrites (*vertical line*), and axons contacting dendrites of post-synaptic neurons (*curved lines*)

work (LGN and other cortical areas in the case of primary visual cortex). See Fig. 3a.

**Networks of Spiking Neurons** The architecture of a network of spiking neurons can be specified by answering the following list of questions:

**How many classes of neurons?** Early studies of networks of spiking neurons focused on one population networks (either excitatory or inhibitory networks). Networks in which there are two populations of neurons—one excitatory, and one inhibitory—have also been analyzed in detail.

**How many neurons in each class?** Analytical studies often use the limit in which the number of neurons goes to infinity (see theoretical tools section below). This is justified by the fact that local networks in the brain are typically composed of a very large number of neurons ( $\sim 10^5$ ). It is also becoming possible to simulate networks of such large numbers of neurons.

**What is the wiring diagram?** Early studies focused on fully connected networks, which are easier to handle analytically. Another popular architecture has been a randomly connected network, with a given connection probability (see Fig. 3b). Such an architecture takes into account the relatively low connection probability of nearby neurons. Relatively few studies of networks of spiking neurons take into account spatial structure, such as a monotonically decaying connection probability.

**What is the nature of the individual couplings between neurons?** Theorists have investigated networks connected by chemical synapses, electrical synapses, or both. More recently, the effect of various types of synaptic plasticity phenomena (both short-term and long-term) has begun to be studied.

### What are the inputs to the neurons in the network?

Popular choices to study the intrinsic dynamics of networks are constant uniform inputs, or noisy inputs, which are uncorrelated from neuron to neuron. To understand the information processing capabilities of networks, one has to choose a model for the type of inputs that are processed by such network. This choice typically depends on the area that is modeled. For example, inputs to a model of a hypercolumn in visual cortex are often taken to represent (fixed or dynamic) oriented bars, and are modeled appropriately.

### Theoretical Tools

Different types of analytical tools are available depending on the nature of the model. Rate models are typically specified by systems of coupled ordinary differential equations. Hence, standard tools can be borrowed from dynamical system theory. One typically starts by investigating the case of constant inputs, for which it is sometimes possible to find fixed points, and the linear stability of these fixed points. Sometimes, it is possible to analyze more complex behaviors, oscillatory states, spatially localized states, waves, etc., and their stability.

Networks of spiking neurons are substantially harder to analyze. Several types of approximations can be used, depending on the strength of coupling and the level of noise: fully connected networks in the weak coupling/weak noise scenario can be studied in the framework of the theory of coupled oscillators. When coupling and noise are strong, it is sometimes possible to approximate synaptic inputs to a neuron as a random Gaussian process, the moments of which depends on both connectivity and activity in the network. For instance, for leaky integrate-and-fire neurons, this leads to the study of the Langevin-like equa-

tion

$$\tau_m \frac{dV}{dt} = -(V - V_L) + \bar{I}_{\text{syn}}(t) + \sigma \sqrt{\tau_m} \xi(t), \quad (6)$$

where  $\bar{I}_{\text{syn}}$  represents the mean synaptic input and  $\xi(t)$  a white noise term representing its fluctuations. In conditions in which correlations between these inputs can be neglected (as when the connection probability between cells is weak), a network of integrate-and-fire neurons can then be characterized by a Fokker–Planck equation that describes the time evolution of the instantaneous distribution of membrane potentials [1,22,105]

$$\tau_m \frac{\partial P}{\partial t} = \frac{\partial}{\partial V} [(V - \bar{I}_{\text{syn}}(t))P] + \frac{\sigma^2}{2} \frac{\partial^2 P}{\partial V^2}, \quad (7)$$

where the mean  $\bar{I}_{\text{syn}}(t)$  is determined self-consistently from the neuron discharges.

In addition to the above mentioned analytical tools, numerical simulations are almost always indispensable, if only to check the validity of the approximations. While simulating rate models is relatively straightforward, simulations of networks of spiking neurons are, again, more demanding. In addition to standard finite-difference integration schemes [53,57,101], it is sometimes possible to use event-driven schemes, in which one jumps from one spike to the next, using an exact integration of network dynamics during inter-spike intervals (see [88]).

### Intrinsic Network Dynamics

We start by a description of intrinsic network dynamics, that is, dynamics in absence of spatially or temporally structured inputs. In rate models, this means taking a constant and uniform input  $I_{\text{ext}}(x, t) = I_{\text{ext}}$ ; in networks of spiking neurons, this means taking either a constant deterministic input, or a noisy stationary input. Noise is typically chosen to be a Gaussian random noise, uncorrelated from neuron to neuron, or uncorrelated Poisson processes independently activating all the neurons of the network.

The goal is then to understand the types of dynamics that can be generated by the network, independently of its inputs. Experimentally, intrinsic dynamics of a network can be observed in vitro, or in vivo, either in anesthetized animals or in awake animals in absence of specific inputs that activate the particular area under observation. Such experiments reveal many different types of activity that modelers have been trying to understand. A prominent and ubiquitous finding in in vivo recordings is irregular background activity at low rates. The issue of the origin and mechanisms of this activity can be addressed using networks of spiking neurons, and will be the topic

of Subsect. “Irregular Firing at Low Rates”. We will then move to the issue of multistability in networks, which can be considered in the simple setting of rate models. Finally, we will discuss the issue of synchronization and oscillations, which can be addressed both with rate models and spiking neuron models.

### Irregular Firing at Low Rates

Recordings of cortical neurons in vivo show very irregular activity. The distributions of intervals between successive action potential emissions, commonly called inter spike intervals (ISI), are approximately Poissonian in a variety of cases. Understanding this irregularity, is a first theoretical challenge [15,111]. Cortical cells receive on the order of  $10^4$  synapses. This large number should tend to promote regular firing since fluctuations of the total input should be small compared to its mean, unless inputs are strongly correlated. Another related difficulty is that measurements of synaptic weights [40,56,87,110] (see [10] for a short survey) indicate that the detectable connections between pyramidal cells are relatively strong with a mean value of 0.5–1 mV, as measured from the peak somatic depolarization that they provoke. As a consequence, thousands of presynaptic pyramidal cells spiking at a low rate of a few Hertz should provide a strong depolarization of the post synaptic cells. It is thus not clear how a low rate of spike emission can be maintained in a recurrently coupled network. One loophole in the argument is that it does not take inhibition into account. Inhibition can potentially solve both puzzles at once when inhibitory inputs balance on average excitatory ones [6,108,120,121]. This drastically reduces the input current so that low firing rates are possible. Moreover, current fluctuations can be comparable to the mean current, allowing for irregular firing.

These studies have shown that irregular firing is obtained in randomly connected networks of spiking neurons, as soon as recurrent inhibition is strong enough. Remarkably, such an irregular firing can be obtained even when the inputs to the network are deterministic, that is, in the absence of noise [120]. The nature of this irregular firing remains the subject of intense study. Irregular firing can be associated with complex but stable trajectories in phase space [63,127], or with chaotic dynamics [127].

Recent data provide some evidence for a balance between inhibition and excitation in cortical networks [109]. Somewhat unexpectedly, a balance of excitation and inhibition has also been found in the rhythmic input onto motor neurons during scratching behavior in turtles [13]. It is proposed that the associated enhanced irregularity can be used to produce a smoother muscle excitation.

### Attractors with Inhomogeneously Distributed Activity

A recurring idea in theoretical studies is that structured connectivity can lead the dynamics of a given neural network to settle in one of several coexisting attractors with mean activity differently distributed among the different neurons of the networks. It is, for instance, central to Hopfield's view of memory storage in neural networks [58], as well as to proposed explanations for persistent activity underlying short-term memory.

Many designs of this type are based on short-range recurrent excitation together with long-range inhibition. The simplest embodies the so-called "winner-take-all" mechanism. It consists of two-recurrently coupled excitatory sub-networks inhibiting each other. This can be described in a rate model framework as

$$\tau \frac{dr_1}{dt} = -r_1 + \Phi(I_0 - Jr_2) \quad (8)$$

$$\tau \frac{dr_2}{dt} = -r_2 + \Phi(I_0 - Jr_1), \quad (9)$$

where  $r_1$  and  $r_2$  describe the activity of each sub-network and  $J > 0$  the magnitude of recurrent inhibition. The homogeneous steady state has  $r_1 = r_2 = r_s$  with

$$r_s = \Phi(I_0 - Jr_s). \quad (10)$$

However, linearization of Eqs. (8), (9) shows that homogeneously distributed activity is unstable when inhibition is sufficiently strong (the bifurcation itself can be supercritical or subcritical depending on the sign of the third derivative of  $\Phi$ ). In this case, the symmetry between the two neuronal populations is broken and there are two possible coexisting attractors with the activity enhanced in either sub-network enhanced and suppressed in the other one.

This scheme can be simply extended to several sub-networks (with global rather than reciprocal inhibition) to enlarge the possible number of coexisting attractors. Some neuronal networks, such as, for instance, the head-direction cell network, are thought to possess a line of attractors instead of a discrete set of attractors. This can again be obtained by generalizing Eqs. (8), (9) to cells labeled by a continuous parameter, which we denote by  $\theta$ , with excitatory connections between cells with close values of  $\theta$  and inhibition between cells with more distant values [5,39,48]. For illustrative purposes, we describe one such simple scheme proposed in [12], where for definiteness the continuous parameter is taken to correspond to an angle. This so-

called "ring model" is a specific case of Eq. (5)

$$\begin{aligned} \tau \frac{d}{dt} r(\theta, t) \\ = -r(\theta, t) + \Phi \left[ I_{\text{ext}} + \int_{-\pi}^{\pi} \frac{d\theta'}{2\pi} J(\theta - \theta') r(\theta', t) \right], \\ -\pi \leq \theta \leq \pi. \end{aligned} \quad (11)$$

It considerably simplifies the analysis [12] to restrict the synaptic coupling function to its first harmonic,  $J = J_0 + J_1 \cos(\theta - \theta')$  and to take  $\Phi$  to be a threshold linear f-I curve,  $\Phi[I] = \beta[I]_+$  that is  $\Phi[I] = \beta[I]$  for  $I \geq T$  and 0 otherwise. Integration over  $\theta$  reduces the dynamics to coupled differential equations for the first three moments  $r_0$ ,  $r_{11}$  and  $r_{12}$  of  $r$  with

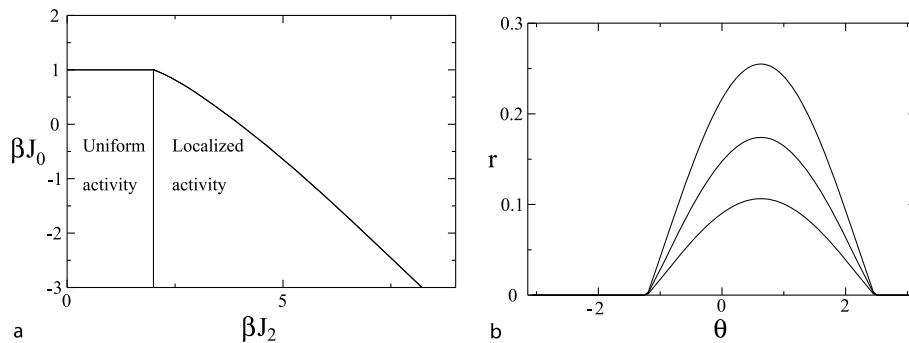
$$\begin{aligned} r_0 &= \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} r(\theta), \\ r_{11} &= \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} r(\theta) \cos(\theta), \\ r_{12} &= \int_{-\pi}^{\pi} \frac{d\theta}{2\pi} r(\theta) \sin(\theta). \end{aligned} \quad (12)$$

The homogeneous activity ( $r(\theta) = r_0$ ) is non-zero for  $I_{\text{ext}} > 0$ ,  $r_0 = I_{\text{ext}}/(1 - \beta J_0)$ ;  $\beta J_0 < 1$  is needed for the dynamics to be well-behaved that is, all-to-all interactions should be inhibitory, or at least not too excitatory, to prevent activity growth without bounds (or saturated activity when saturation is included in  $\Phi$ ). However, homogeneous activity is unstable when interaction between cells with close values of  $\theta$  is large enough that is, here for  $\beta J_1 > 2$ . Then, the network attractor is formed by a line of solutions in the form

$$r(\theta) = A[\cos(\theta - \theta_0) - \cos(\theta_c)]_+, \quad -\pi \leq \theta_0 \leq \pi. \quad (13)$$

Thus, the network activity is confined to cells with parameter  $\theta$  in a interval of values of width  $2\theta_c$  around an arbitrary angle  $\theta_0$ . The constant  $\theta_c$ , the solution selectivity, depends only on the magnitude of  $\beta J_1$ . Localized states of activity and their domain of existence in parameter space are shown in Fig. 4.

The ring model is a nice example of a model with a continuous line of attractors but it also exemplifies difficulties inherent to this type of models. The main problem is that a line of attractors is a fragile structure. It is not resistant to perturbations coming, for instance, from heterogeneity in the connection strengths or neurons' properties. If this design is used in real neural systems, there should exist additional mechanisms that maintain the line of attractors and avoid its breaking to a discrete set. Some proposals exist [70,102], but the existence of a line attractor in an actual neural network remains to be demonstrated.



Neuronal Dynamics, Figure 4

Ring model. **a** Phase diagram showing the position of the uniform and localized activity for the model of Eq. (11). **b** Profiles of localized activity for  $J_{\text{ext}} = .1$ ,  $J_2 = 3$ . and  $J_0 = -2, -1, . - 5$  (from bottom to top). Adapted from [12]

### Dynamics at Various Time Scales

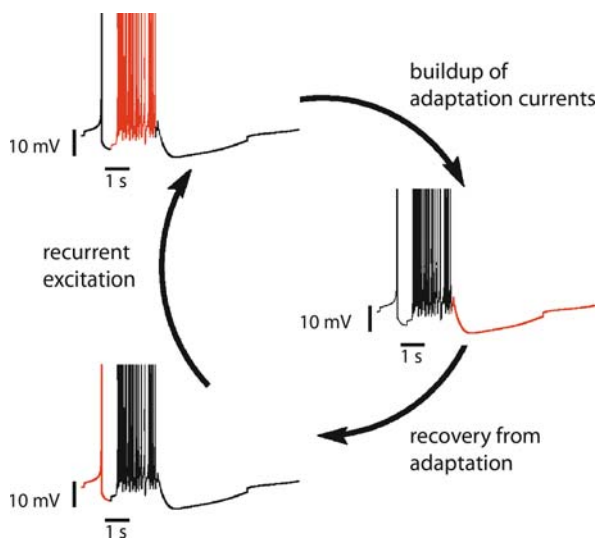
As has already been stressed, spontaneous activity in a variety of neural systems exhibits rhythmicity at time scales that range from a few millisecond to many seconds. We restrict ourselves to rhythms which are intrinsically of electrophysiological origins (unlike for instance, circadian rhythms which are accompanied by changes of neuronal excitability but which are thought to be primarily driven by changes in gene expression).

**Bursts of Activity** At the lower end of the spectrum, networks which produce repeated bursts of activities at a time scale of a few tenths of seconds to a few seconds are observed in different cases. They seem to be an essential component of pacemakers generating locomotion [29] or respiration [69]. Repeated bursts of activity are also observed in neuron cultures [81,93,96]. An important additional example is given by the observation of up-down states in cortical slices [107], which are proposed to be analogous to slow oscillations during slow-wave sleep.

Specific models have been proposed for different systems but they appear to be based on the same general ideas. First, recurrent excitation allows for the self-sustaining bursting state. Second, this active state terminates when a slower building process has reached sufficient magnitude to prevent self-sustainment of the active state. The main contenders for the slow process are synaptic depression (see Subsect. “Dynamic Synapses”) or some intracellular mechanism such as a slow increase of an ionic concentration that opens an ion-gated channel once it has reached a threshold level. The generation of bursts of activity by synaptic depression has been modeled at a general theoretical level in [118]. In the context of respiratory rhythm generation, burst termination has been modeled as due to the slow inactivation of a  $\text{Na}^+$  current (that is, termination

of a depolarizing current) or by the activation of a calcium-gated potassium (that is, hyperpolarizing) current. For up-down states, a sodium-dependent potassium conductance has been hypothesized to mediate their termination and has been taken into account in a proposed model [32]. The mechanism is illustrated in Fig. 5. Both calcium and sodium-gated potassium conductances have been identified in spinal neurons [123]. It is worth noting however that in most cases the actual mechanism at work has not been experimentally pinpointed.

Besides burst termination, the cause of a burst should be determined to see how a repetition at about 1 Hz can



Neuronal Dynamics, Figure 5

Mechanism of slow oscillations in cortical slices. Recurrent excitation provokes a burst of activity which is terminated by a slow adaptation current. Activity resumes when the adaptation current has decayed. From [32]



come about. The general mechanisms appear again to be of two different types. Slow recovery from the termination process (for instance, return of concentration to a low level by the action of an ionic pump) can deterministically bring the networks above a threshold for excitation. Alternately, recovery can lead the network in a low activity state from which it can stochastically transit to the active state [126].

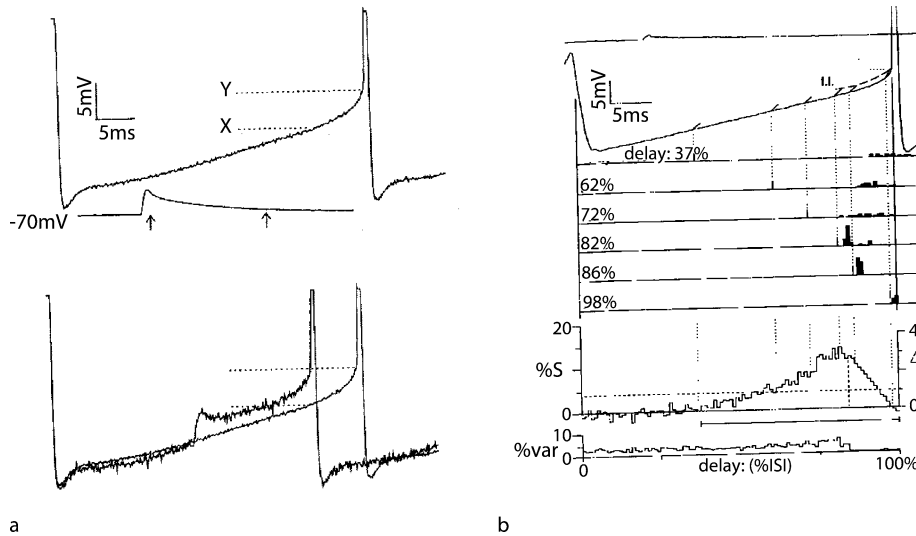
It is worth noting here that both the size distribution and duration distribution of bursts of activity in cortical slices have been examined using multi-electrode arrays [11]. Power law distributions corresponding to critical branching processes, reminiscent of a self-organized critical phenomenon [9], have been found. This was shown to emerge naturally from synaptic depression in a simplified IF neuron model without leak [77]. The applicability of this result to more physiologically realistic networks remains to be examined.

**Wandering Between Different Attractors** Bursts of activity can be seen as a succession of jumps between two attractors with homogeneously distributed activity—a high activity state and a low activity state—mediated by a slow process. With structured connectivity, similar mechanisms can mediate transitions between a larger number of more complex coexisting attractors, as those described in Subsect. “[Attractors with Inhomogeneously Distributed Activity](#)”. A given active state then consists of one particular subnetwork of active neurons among several possible ones. Some evidence for such stochastic visits of coexisting attractors has been provided both in vitro and in vivo. Using calcium imaging, mouse neocortical slices have been found to display sparse spontaneous activity and sets of coactive neurons that repeatedly appear above chance level [33]. In vivo, neurons of cat visual cortex responding to the same preferred orientation (the so-called orientation maps) have also been reported to be spontaneously coactive, with sets belonging to different orientations appearing in time in a stochastic manner [67,115]. A model with biologically plausible connectivity and attractors consisting of orientation maps is able to reproduce these remarkable data [17]. It should be noted, however, that the experimental results seem also to be interpretable as fluctuations about a single background state driven by correlated thalamic inputs [49].

Interestingly, a similar set of ideas has been used to model slow oscillations in higher cognitive tasks. One interesting example is binocular rivalry: when the right and left eyes are presented with two different images, the perceived pattern alternates between the two images every few seconds. Moreover, several functional magnetic resonance imaging studies have shown that fluctuation in neu-

ral activity in several brain areas correlates with perception changes [80,100,114]. The dominance of the perception of one image over the other one has been modeled as a competition between two recurrently coupled excitatory networks reciprocally inhibiting each other, the “winner-take-all” design described in Subsect. “[Attractors with Inhomogeneously Distributed Activity](#)”. Without further elaboration this gives rise to two attractors: either one of the two networks can be active with the other silenced. Similarly to burst or upstate termination, alternation between attractors (and percepts) can arise by supplementing this basic design with a slow process. It has been modeled either as coming in a deterministic way from synaptic depression [124] or from a stochastic jump from one of the two attractors to the other [92].

**Synchronization of Periodically Firing Neurons** Neural recordings shows oscillations in diverse frequency bands. They depend both on the neural structure and, when recorded in vivo, on the type of activity that the animal is performing. These rhythms, which are seen in EEG or local field potential, emerge from the coordinated activity of many neurons. Synchronization of nonlinear oscillators [99] has of course been thoroughly studied since its discovery by Huyghens. Some of the tools developed [73] have been directly applied to neurons in a regime where they emit action potentials in a periodic manner [52,122]. Many experimental and theoretical studies have focused on single neurons and coupled pairs of neurons to infer synchronization properties of large networks of weakly coupled neurons. This has acquired renewed importance with the discovery that inhibitory interneurons in the same subtype category appear to be preferentially coupled [45,47]. For weak coupling, the dynamics of an oscillator can be described by the behavior of its phase, that is, its position along its limit cycle. How the spike advance/delay produced by a short current injection depends on the time of injection during the interspike interval is a crucial quantity for synchronization. This so-called phase-response-curve (PRC) has been studied in different models [52,122] with the result that generally a small and short depolarization advances the next spike, and does so to a greater extent the more closely it occurs before spike emission. This is referred to as a type I PRC. In some cases, though, (as in the original Hodgkin-Huxley model) a less intuitive type II PRC is found: a small depolarization just after the spike has a greater effect on hyperpolarizing currents than on depolarizing ones and results in a delay of the next spike. The PRC has also been experimentally measured for neocortical neurons. The results of the early experiment of [103] are shown in Fig. 6



Neuronal Dynamics, Figure 6

Measurement of phase-response curve (PRC) for neocortical neurons [103]. The authors injected a short depolarizing current pulse at different times during the interspike interval. As shown on the *left panel*, this results in a timing change for the spike following the current injection. The *right panel* show results for different injection times as well as the PRC which summarizes the data by giving the spike time changes as function of the current injection time (expressed here as a percentage of the ISI). Adapted from [103]

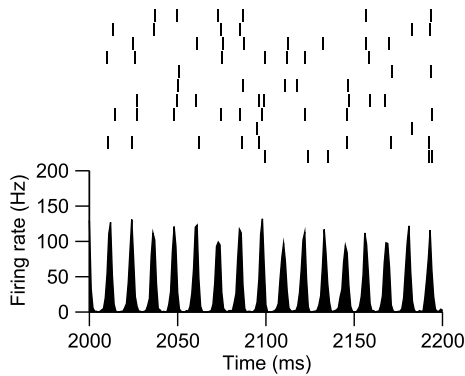
and give a type I PRC. For instantaneous synaptic transmission, this would lead one to expect that excitatory interactions tend to make two identical neurons spike in phase while inhibitory ones would push them to antiphase spiking. It was therefore an important theoretical finding that, for more realistic synaptic currents with finite rise and decay, this is not generally the case and that inhibition can be more efficient than excitation for synchronization. For excitatory synapses, complete in-phase spiking is generally unstable for type I PRC and a finite dephasing exists even for identical neurons. On the contrary for inhibitory synapses, in-phase and anti-phase spiking are both possible at low frequency, a bistable situation. However the attraction basin of the in-phase state grows with frequency, and above a threshold frequency the anti-phase state disappears. This leads one to expect a transition from antiphase to in-phase spiking for a coupled pair of identical interneurons as their firing frequency increases. This has been verified experimentally [46] as well as other similar predictions for gap junctions or on the combined effect of inhibitory synapses and gap junctions [31,78,84].

**Sparsely Synchronized Oscillations** Although a substantial amount of work has been devoted to analyzing periodically firing neurons with a minimal amount of noise and heterogeneity, this seems to be a rather infrequent situation in the brain. Strong heterogeneity appears to be the norm rather than the exception. Moreover, as explained in Subject. “Irregular Firing at Low Rates,” the ac-

tivity of most neurons appears quite irregular and their discharge rate is often low compared to the frequency of gamma oscillations and faster rhythms. This is also the case in vitro when pharmacologically-induced activity in hippocampal slices results in spontaneous gamma oscillations [41]. Modeling studies have shown that a mechanism distinct from the Huyghens-type synchronization of oscillators can produce a fast and robust rhythm at the network level with neurons discharging irregularly and at a lower frequency [23]. An example from a simulation of a network of IF neurons is shown in Fig. 7. Recurrent inhibition again plays an important role in these sparsely synchronized oscillations. The basic idea can be understood from a rate model description similar to Eq. (1)

$$\tau \frac{dr}{dt} = -r + \Phi[I_{\text{ext}} + Jr(t - D)], \quad (14)$$

where the activity  $r(t)$  of the network at time  $t$  is influenced by its activity at a previous time  $t - D$ . The delay  $D$  models in a crude but effective way the latency and finite kinetics of synaptic currents (see Eq. (4)). It is not difficult to show that for  $J < J_c < 0$ , that is, for a sufficiently strong recurrent inhibition, the activity  $r(t)$  is attracted to a limit cycle and oscillates periodically. The oscillation frequency is of order  $1/D$  and increases from  $f = 1/(4D)$  for delays short compared to  $\tau$  to  $f = 1/(2D)$  for delays long compared to  $\tau$ . It is simple to see the origin of the oscillations: an increase of activity in the network at time  $t$  provokes an increase of recurrent inhibition. This results in



**Neuronal Dynamics, Figure 7**

Oscillations with sparsely firing neurons in a fully connected network of 1000 leaky integrate-and-fire neurons receiving independent white noise sources. The *top panel* shows a raster of 10 neurons, while the *bottom panel* shows the network instantaneous firing rate (computed in 1 ms bins). The network oscillates at about 90 Hz, while single cells fire at about 30 Hz, as predicted by theory [22,25] (synaptic time constants: 1 ms latency, 1 ms rise time, 6 ms decay time). Reproduced from [21]

a decrease in network activity at about  $t + D$ , since transmission from one neuron to the next takes a time  $D$ . This decrease of activity at  $t + D$  itself again generates an increase of activity at about  $t + 2D$  from which the cycle can continue. One sees that the frequency of the network oscillation, of order  $1/D$ , is directly linked to the kinetics of synaptic transmission but not related to the neuron discharge rate  $r$ . While the rate description of Eq. (14) simply captures some characteristics of network oscillations in the sparsely synchronized regimes, it is far from providing accurate estimates, in particular for the oscillation threshold. A more complex mathematical description based on the Fokker-Planck equation has been developed to this end [22,25]. It allows, moreover, the inclusion of various features of single neuron dynamics such as the finite rise time of the action potential [42] or of an eventual resonance in subthreshold dynamics [26,104] (see [23] for a short review).

Although many experimental recordings are suggestive of this kind of oscillations in the brain, it is not easy to eliminate the possibility that the observed rhythm is created by a population of fast-spiking unidentified cells. At present, two of the best studied examples are provided by fast oscillations in the hippocampus and the cerebellum. The “ultrafast” ripple oscillations (140–200 Hz) are the fastest among the numerous hippocampal rhythms. They occur together with “sharp waves” during awake immobility and slow wave sleep in rats. Recordings show that both the discharge rates of pyramidal cells and interneu-

rons (average rates 8 Hz and 30 Hz respectively [34]) are much lower than the frequency of the population oscillation. In the cerebellum, fast oscillations were discovered by Adrian in one of the first intracranial recordings [3]. More recently, de Solages et al. [37] have provided strong evidence that these fast (150–250 Hz) oscillations are emerging from recurrent inhibition among Purkinje cells that themselves fire at an average of about 40 Hz.

**Wave Propagation** The propagation of waves of neural activity is certainly an important topic but one that has traditionally been difficult to study experimentally. It has therefore been the subject of relatively few specific theoretical works. Wave propagation in neural networks has been classically analyzed in the framework of rate models [5,39] with results for wave existence similar to those for general excitable media (for which many studies exist motivated, for instance, by applications to chemical waves in the Belousov-Zhabotinsky reaction or, in a biological context, to electrical waves in cardiac tissue). Advances in imaging techniques are now providing more data on neural activity propagation, in slices [18,72,107], in cell cultures [62] and even in vivo [98]. In parallel, theoretical interest in spiking models is developing and results have been obtained either by the analysis of simplified cases [50] or by simulations [32]. The results highlight in particular the role played by inhibition and the potential importance of long-range connections in wave propagation.

### Stimulus Driven Dynamics

What happens when external stimuli are presented to a network of neurons? An external stimulus is usually modeled as an increase (or decrease) of external inputs to specific groups of neurons. For example, in the ring model, presentation of an external stimulus representing an oriented bar during an interval  $[0, T]$  is described by  $I_{\text{ext}}(\theta, t) = I_0 + I_1 \cos(\theta - \theta_s) \Theta(t) \Theta(T - t)$ , where  $\theta_s$  is the angle of the presented bar. In a spiking neuron model, an external stimulus can be represented by a transient change of the mean currents impinging on a subset of neurons, or by a transient change in the frequency of incoming spike trains. Conceptually, these transient inputs can lead to three types of phenomena: (i) During the transient input, the nature of the dynamics of the network can change qualitatively. For example, the network might switch from an asynchronous to a synchronous state or it might switch from a uniform to a spatially localized state. (ii) In multistable networks, the stimulus can switch the network from one state to another, and the network will then stay in this particular state, thereby maintaining a short-term mem-

ory of the stimulus that was presented. (iii) Presentation of a stimulus can induce synaptic plasticity between neurons that are activated/inactivated by the stimulus, leading to a remodeling of the network, and therefore potentially to a change in the repertoire of states that the network dynamics can sustain. We now briefly examine these three phenomena before focusing on specific systems.

### Dynamics During Stimulus Presentation

**Selective Amplification** The simplest change induced by external stimuli is a change in firing rates of the neurons. This change in rate is governed by external inputs, but also by recurrent connectivity. This can be best studied in the context of rate models (see [35]). In general, excitatory networks tend to amplify external inputs, while inhibitory networks tend to attenuate them. With spatially structured connectivity, the network will tend to amplify certain inputs, at the same time attenuating another. A typical example is represented by the ring model of Subsect. “[Attractors with Inhomogeneously Distributed Activity](#)” with a ‘Mexican-hat’ connectivity – with this type of connectivity (excitatory at short-range, inhibitory at long-range), the network tends to amplify spatially localized inputs, while it will attenuate inputs with no spatial selectivity. Thus, the network has the property of selective amplification. When the excitatory connectivity is strong enough and one enters the marginal phase, the network generates spatially localized states even in absence of external inputs. In this case, the external stimulus selects one of the possible network states (a state in which the activity is peaked around the most strongly activated neuron) out of the repertoire of possible states (a continuum in the case of the ring model). This property is consistent with several experimental findings in various areas [67,79].

### Switching from Asynchronous to Synchronous State

Another effect of external stimuli can be to change the degree of synchrony in the network. For example, a stimulus can switch a network from an asynchronous state to a synchronous state (or vice versa). This is typically what happens in randomly connected networks of excitatory and inhibitory neurons (or purely inhibitory networks) when inhibition is stronger than excitation [20,22]. This is due to the fact that external inputs tend to increase the average firing rate of the network, leading to an effective increase in the strength of inhibitory feedback. This increase can potentially make the network switch from the asynchronous to the synchronous region.

This emergence of synchronous activity induced by external inputs is reminiscent of many experimental find-

ings, from oscillations induced by odors in the olfactory system (to be discussed in more detail below), to oscillations correlated with selective visual attention in the visual system of the primate [44]. The emergence of oscillations in the presence of visual inputs has also been hypothesized to allow the system to solve the so-called ‘binding problem,’ though this issue remains hotly debated [59,71,106].

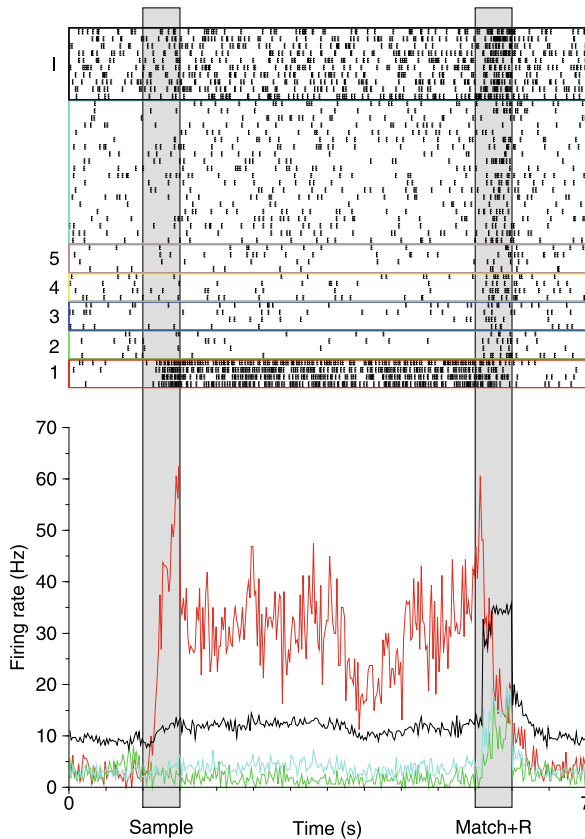
### Multistability and Working Memory

In multistable networks, stimuli can potentially switch the network from one state to another. For example, in winner-take-all rate models, a stimulus will switch the network to an attractor state in which the population that received the largest input is active, while all others are inactive. The fact that this network configuration is an attractor means that the network maintains, in short-term memory, some information about which stimulus it was shown. This property is the hallmark of associative memory models, where many ‘memory states’ in which subpopulations of neurons are selectively active at higher rates, are attractors of the network dynamics, thanks to strong excitatory feedback inside such sub-populations. It is widely believed to form the basic mechanism underlying working memory, that is, the active maintenance of information in short-term memory. Such networks have been investigated extensively in the last three decades, from networks of simplified binary neurons [58] to networks of spiking neurons of increasing realism see [6,24]). The working of such a network is illustrated in Fig. 8.

### Effect of Long-Term Synaptic Plasticity on Network Dynamics

Another potential effect of external stimuli is to modify the synaptic structure of the network. For example, in associative memory models, external stimuli are assumed to lead to synaptic modifications induced by increases or decreases of firing rates of single neurons. Hebbian learning rules posit that synapses connecting two neurons which are strongly activated by the same stimulus will potentiate, while synapses connecting one strongly activated neuron to an inactivated neuron will depress. This learning dynamic tends to create a synaptic structure that leads to attractors strongly correlated with the state of the network during stimulus presentation.

Most network studies have separated neuronal and synaptic dynamics, assuming much slower synaptic dynamics. Typically, one studies the network with a fixed synaptic structure that incorporates, in a Hebbian way, learning of patterns of activity presented to the network in the past. More recently, there have been stud-



**Neuronal Dynamics, Figure 8**

Dynamics of a randomly connected network of excitatory and inhibitory neurons with working memory properties (*top*: raster; *bottom*: average firing rate of different populations of neurons). Inhibition is strong enough to stabilize the network in a background state (before ‘sample’ presentation). External stimuli have been ‘stored’ in the synaptic matrix, through enhanced synaptic connectivity between neurons that belong to a sub-population that is activated by a stimulus (sub-populations 1–5 in the figure). Presentation of an external stimulus (sample 1 in the figure) increases the firing rate of the relevant neurons (see *red curve* in *bottom panel*, average firing rate of neurons in population 1). This enhanced firing rate survives removal of the stimulus because of the strong excitatory connectivity between 1 neurons. Hence, the stimulus has switched the network from the background state to the memory state corresponding to stimulus 1. Finally, a sufficiently strong non-specific stimulus switches the network back to the background state. From [24]

ies of networks with double dynamics of neurons and synapses [60,91].

### Dynamics and Information Processing: Two Specific Examples

In some neural structures, the role of dynamics in information processing has been particularly scrutinized. We

briefly describe here two prominent cases and some views and hypotheses that they have suggested.

**Dynamics and Odor Discrimination** As recalled previously, it has been known since the recordings of Adrian [4] that odors promote gamma oscillations in the olfactory bulb. The role of oscillations in odor processing has, however, remained unclear in spite of interesting proposals [43]. Olfaction is an evolutionarily ancient function and the associated neural structure are analogous in very different species. Recent recordings in insects are shedding new light on the role of dynamics in odor discrimination. The antennal lobe in insects corresponds to the olfactory bulb in mammals. It receives input from neurons in the olfactory epithelium and its principal neurons project to the next structure, the “mushroom body”. In the locust, puffs of two close odors induce antennal lobe activities that are similar at start (as measured from the activity of a sample of the eight hundred projection neurons) but that diverge in a few hundred milliseconds [113]. The antennal lobe oscillations shape its output as successive volleys of projection neuron spikes emitted at a rate of about 20–30 Hz. The numerous Kenyon cells in the mushroom body then appear to function as coincidence detectors that process each projection neuron spike volley independently of the others [89,113] (reset between each volley being provided by non-specific feedforward inhibition coming from lateral horn interneurons [97]). The precise timing of the whole process appears to allow STDP plasticity at the Kenyon cells output synapses [30] and presumably memory formation in a way that remains to be related to behavioral experiments and genetic data [54,66].

**Place Maps and Oscillations in the Hippocampus** The hippocampus is another brain region in which oscillations are prominent and one in which their role in information processing is among the best studied. The hippocampus is an important structure for navigation. “Place” cells that fire at particular locations have been discovered thirty years ago in the rat hippocampus [94]. Different place cells discharge in different “place fields”, of about 25 cm in size, and recording several cells allows the determination of the animal position in its environment. However, place cell firing is also strongly modulated at theta frequency. The phase of a place cell discharge advances as the animal enters the place cell field [95]. So, monitoring precise spike timing with respect to the theta oscillations provides supplementary information and a more accurate position determination [64,125]. Several models have been proposed for place cell formation during exploration of a new en-



vironment, some based on network dynamics [116] related to that exposed in Subsect. “[Attractors with Inhomogeneously Distributed Activity](#)”, other based on temporal modulation of single cell spiking created by addition of oscillations at different frequencies [65,76,95]. Supplementary complexity and theoretical puzzle has been added by the startling discovery of “grid” cells [51,90] in the entorhinal cortex, a neural area in the hippocampal formation by which transits much of the sensory information that reaches the hippocampus.

### Future Directions

We are still far from having a precise understanding of how neural systems operate, but there are daily advances at all levels from the description of detailed molecular mechanisms to that of high-level cognitive processes. We limit ourselves here to underlining some areas where further investigations of dynamics are clearly needed.

We have focused on networks with very little spatial structure but propagation and spatio-temporal dynamics most surely play important roles in neural information processing. This appears to be a very rich domain that imaging and theoretical investigations will help to explore.

A related question is that of the coordination between different neural areas. How do dynamics in one structure, for instance oscillations, serve to transmit and process information in a connected structure? As mentioned above, recent investigations of olfaction in insects provide a fascinating glimpse on this question. It will clearly be necessary to address it more generally in spite of the experimental difficulties.

Our emphasis in this article has been on electrical activity and its transmission. Another all-important aspect is neuromodulation which affects neuron properties on a slower time scale and probably a less local spatial scale. At present, the relation between these two facets of neural activity has been considered in few theoretical studies, but this needs to be thoroughly addressed in future work.

Homeostasis, or how dynamical properties of neuronal networks are maintained, is another question that needs analysis. Interesting experimental and theoretical work has been performed using the stomatogastric of the lobster as an example [75,85], but certainly this important topic requires further scrutiny.

These few questions among many others should make it clear that investigation of neuronal dynamics requires the development of powerful experimental techniques and theoretical analyses and that the subject will remain a topic of intense research in the forthcoming years.

## Bibliography

### Primary Literature

1. Abbott LF, van Vreeswijk C (1993) Asynchronous states in a network of pulse-coupled oscillators. *Phys Rev E* 48:1483–1490
2. Abbott LF, Varela JA, Sen K, Nelson SB (1997) Synaptic depression and cortical gain control. *Science* 275:220–224
3. Adrian ED (1934) Discharge frequencies in the cerebral and cerebellar cortex. *Proc Physiol Soc* 83:32–33
4. Adrian ED (1942) Olfactory reactions in the brain of the hedgehog. *J Physiol* 100:459–473
5. Amari S (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol Cybern* 27:77–87
6. Amit DJ, Brunel N (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb Cortex* 7:237–252
7. Amit DJ, Tsodyks MV (1991) Quantitative study of attractor neural network retrieving at low spike rates I: Substrate – spikes, rates and neuronal gain. *Network* 2:259–274
8. Badel L, Lefort S, Brette R, Petersen CCH, Gerstner W, Richardson MJE (2008) Dynamic I-V curves are reliable predictors of naturalistic pyramidal-neuron voltage traces. *J Neurophysiol* 99:656–666
9. Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. *Phys Rev A* 38:364–374
10. Barbour B, Brunel N, Hakim V, Nadal J (2007) What can we learn from synaptic weight distributions? *Trends Neurosci* 30:622–629
11. Beggs JM, Plenz D (2003) Neuronal avalanches in neocortical circuits. *J Neurosci* 23:11167–77
12. Ben-Yishai R, Bar-Or RL, Sompolinsky H (1995) Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci USA* 92:3844–3848
13. Berg RW, Alaburda A, Hounsgaard J (2007) Balanced inhibition and excitation drive spike activity in spinal half center. *Science* 315:390–3
14. Berger H (1929) Über das Elektroenkephalogramm des Menschen. *Arch Psychiatr Nervenkrankh* 87:527–570
15. Bernander O, Douglas RJ, Martin KA, Koch C (1991) Synaptic background activity determines spatio-temporal integration in single pyramidal cells. *Proc Natl Acad Sci USA* 88:11569–11573
16. Bliss TVP, Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361:31–39
17. Blumenfeld B, Bibitchkov D, Tsodyks MV (2006) Neural network model of the primary visual cortex: From functional architecture to lateral connectivity and back. *J Comput Neurosci* 20:219–241
18. Bolea S, Sanchez-Andres J, Huang X, Wu J (2006) Initiation and propagation of neuronal coactivation in the developing hippocampus. *J Neurophysiol* 95:552–561
19. Bromfield EB, Cavazos JE, Sirven JI (2006) An introduction to Epilepsy. American Epilepsy Society, Bethesda
20. Brunel N (2000) Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J Comput Neurosci* 8:183–208
21. Brunel N (2008) Modeling point neurons: from Hodgkin–Huxley to Integrate-and-Fire. In: De Schutter (ed) *Computa-*

- tional modeling methods for neuroscientists. Mit Press, Cambridge
22. Brunel N, Hakim V (1999) Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Comp* 11:1621–1671
  23. Brunel N, Hakim V (2008) Sparsely synchronized neuronal oscillations. *Chaos* 18:015113
  24. Brunel N, Wang X-J (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J Comput Neurosci* 11:63–85
  25. Brunel N, Wang X-J (2003) What determines the frequency of fast network oscillations with irregular neural discharges? *J Neurophysiol* 90:415–430
  26. Brunel N, Hakim V, Richardson MJE (2003) Firing rate resonance in a generalized integrate-and-fire neuron with sub-threshold resonance. *Phys Rev E* 67:051916
  27. Burkitt AN (2006) A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biol Cybern* 95:1–19
  28. Burkitt AN (2006) A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties. *Biol Cybern* 95:97–112
  29. Cangiano L, Grillner S (2005) Mechanisms of rhythm generation in a spinal network deprived of crossed connections: the lamprey hemichord. *J Neurosci* 25:923–35
  30. Cassenauer J, Laurent G (2007) Hebbian stdp in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. *Nature* 448:709–713
  31. Chow CC, Kopell N (2000) Dynamics of spiking neurons with electrical coupling. *Neural Comput* 12:1643–78
  32. Compte A, Sanchez-Vives MV, McCormick DA, Wang X-J (2003) Cellular and network mechanisms of slow oscillatory activity (<1 Hz) and wave propagations in a cortical network model. *J Neurophysiol* 90:2707–2725
  33. Cossart R, Aronov D, Yuste R (2003) Attractor dynamics of network up states in the neocortex. *Nature* 423:283–288
  34. Csicsvari J, Hirase H, Czurko A, Buzsáki G (1998) Reliability and state dependence of pyramidal cell-interneuron synapses in the hippocampus: an ensemble approach in the behaving rat. *Neuron* 21:179–189
  35. Dayan P, Abbott L (2001) Theoretical neuroscience. MIT Press, Cambridge
  36. De Schutter E (1999) Using realistic models to study synaptic integration in cerebellar Purkinje cells. *Rev Neurosci* 10:233–245
  37. de Solages C, Szapiro G, Brunel N, Hakim V, Isope P, Buisseret P, Rousseau C, Barbour B, Léna C (2008) High-frequency organization and synchrony of activity in the purkinje cell layer of the cerebellum. *Neuron* 58:775–788
  38. Destexhe A, Mainen ZF, Sejnowski TJ (1998) Kinetic models of synaptic transmission. In: Koch C, Segev I (eds) *Methods in Neuronal Modeling*, 2nd edn. Cambridge, MIT Press, pp 1–26
  39. Ermentrout GB (1998) Neural networks as spatio-temporal pattern-forming systems. *Rep Prog Phys* 61:353–430
  40. Feldmeyer D, Lübke J, Sakmann B (2006) Efficacy and connectivity of intracolumnar pairs of layer 2/3 pyramidal cells in the barrel cortex of juvenile rats. *J Physiol* 575:583–602
  41. Fisahn A, Pike FG, Buhl EH, Paulsen O (1998) Cholinergic induction of network oscillations at 40 Hz in the in vitro. *Nature* 394:186–189
  42. Fourcaud-Trocmé N, Hansel D, van Vreeswijk C, Brunel N (2003) How spike generation mechanisms determine the neuronal response to fluctuating inputs. *J Neurosci* 23:11628–11640
  43. Freeman W (1991) The physiology of perception. *Sci Am* 264:78–85
  44. Fries P, Reynolds J, Rorie A, Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291:1560–1563
  45. Galarreta M, Hestrin S (1999) A network of fast-spiking cells in the neocortex connected by electrical synapses. *Nature* 402(6757):72–5
  46. Gibson JR, Beierlein M, Connors BW (1999) Functional properties of electrical synapses between inhibitory interneurons of neocortical layer 4. *J Neurophysiol* 93:467–80
  47. Gibson JR, Beierlein M, Connors BW (1999) Two networks of electrically coupled inhibitory neurons in neocortex. *Nature* 402(6757):75–9
  48. Gierer A, Meinhardt H (1972) A theory of biological pattern formation. *Kybernetik* 12:30–9
  49. Goldberg JA, Rokni U, Sompolinsky H (2004) Patterns of ongoing activity and the functional architecture of the primary visual cortex. *Neuron* 42:489–500
  50. Golomb D, Ermentrout GB (2002) Slow excitation supports propagation of slow pulses in networks of excitatory and inhibitory populations. *Phys Rev E* 65:061911
  51. Hafting T, Fyhn M, Molden S, Moser M, Moser E (2005) Microstructure of a spatial map in the entorhinal cortex. *Nature* 436:801–806
  52. Hansel D, Mato G, Meunier C (1995) Synchrony in excitatory neural networks. *Neural Computation* 7:307–337
  53. Hansel D, Mato G, Meunier C, Neltner L (1998) On numerical simulations of integrate-and-fire neural networks. *Neural Computation* 10:467–483
  54. Heisenberg M (2003) Mushroom body memoir: from maps to models. *Nat Rev Neurosci* 4:266–275
  55. Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conductance and excitation in nerve. *J Physiol* 117:500–544
  56. Holmgren C, Harkany T, Svennenfors B, Zilberter Y (2003) Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *J Physiol* 551:139–153
  57. Honeycutt RL (1992) Stochastic Runge–Kutta algorithms. I. White noise. *Phys Rev A* 45:600–603
  58. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79:2554–2558
  59. Issue S (1999) Special issue: The binding problem. *Neuron* 24:7–125
  60. Izhikevich E, Gally J, Edelman G (2004) Spike-timing dynamics of neuronal groups. *Cereb Cortex* 14:933–944
  61. Izhikevich EM (2001) Resonate-and-fire neurons. *Neural Networks* 14:883–894
  62. Jacobi S, Moses E (2007) Variability and corresponding amplitude-velocity relation of activity propagating in one-dimensional neural cultures. *J Neurophysiol* 97:3597–3606
  63. Jahnke S, Memmesheimer R-M, Timme M (2008) Stable irregular dynamics in spiking neural networks. *Phys Rev Lett* 100:048102

64. Jensen O, Lisman J (2000) Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. *J Neurophysiol* 83:2602–2609
65. Kamondi A, Acsády L, Wang X-J, Buzsáki G (1998) Theta oscillations in somata and dendrites of hippocampal pyramidal cells in vivo: activity-dependent phase-precession of action potentials. *Hippocampus* 8:244–261
66. Keene A, Waddell S (2007) Drosophila olfactory memory: single genes to complex neural circuits. *Nat Rev Neurosci* 8:341–354
67. Kenet T, Bibitchkov D, Tsodyks MV, Grinvald A, Arieli A (2003) Spontaneously emerging cortical representations of visual attributes. *Nature* 425:954–956
68. Knight BW (1972) Dynamics of encoding in a population of neurons. *J Gen Physiol* 59:734–766
69. Koshiya N, Smith JC (1999) Neuronal pacemaker for breathing visualized in vitro. *Nature* 400:360–63
70. Koulakov AA, Raghavachari S, Kepecs A, Lisman JE (2002) Model for a robust neural integrator. *Nat Neurosci* 5:775–782
71. Kreiter AK, Singer W (1996) Stimulus dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *J Neurosci* 16:2381–2396
72. Kubota D, Colgin L, Casale M, Brucher F, Lynch G (2003) Endogenous waves in hippocampal slices. *J Neurophysiol* 89:81–89
73. Kuramoto Y (1984) Chemical oscillations, waves and turbulence. Springer, New York
74. Lapique L (1907) Recherches quantitatives sur l'excitabilité électrique des nerfs traitée comme une polarisation. *J Physiol Pathol Gen* 9:620–635
75. LeMasson G, Marder E, Abbott L (1993) Activity-dependent regulation of conductances in model neurons. *Science* 259:1915–1917
76. Lengyel M, Szatmáry Z, Erdi P (2003) Dynamically detuned oscillations account for the coupled rate and temporal code of place cell firing. *Hippocampus* 13:700–714
77. Levina A, Hermann JM, Geisel T (2007) Dynamical synapses causing self-organized criticality in neural networks. *Nature Physics* 3:857–860
78. Lewis T, Rinzel J (2003) Dynamics of spiking neurons connected by both inhibitory and electrical coupling. *J Comput Neurosci* 14:283–309
79. Luczak A, Barthó P, Marguet S, Buzsáki G, Harris K (2007) Sequential structure of neocortical spontaneous activity in vivo. *Proc Natl Acad Sci USA* 104:347–352
80. Lumer ED, Friston KJ, Rees G (1998) Neural correlates of perceptual rivalry in the human brain. *Science* 280:1930–1934
81. Maeda E, Robinson HPC, Kawana A (1995) The mechanisms of generation and propagation of synchronized bursting in developing networks of cortical neurons. *J Neurosci* 15:6834–45
82. Malenka RC, Bear MF (2004) LTP and LTD: an embarrassment of riches. *Neuron* 44:5–21
83. Malinow R, Mainen ZF, Hayashi Y (2000) LTP mechanisms: from silence to four-lane traffic. *Curr Opin Neurobiol* 10:352–357
84. Mancilla JG, Lewis TJ, Pinto DJ, Rinzel J, Connors BW (2007) Synchronization of electrically coupled pairs of inhibitory interneurons in neocortex. *J Neurosci* 27:2058–73
85. Marder E, Goaillard J (2006) Variability, compensation and homeostasis in neuron and network function. *Nat Rev Neurosci* 7:563–574
86. Markram H (2006) The blue brain project. *Nat Rev Neurosci* 7:153–160
87. Mason A, Nicoll A, Stratford K (1991) Synaptic transmission between individual pyramidal neurons of the rat visual cortex in vitro. *J Neurosci* 11:72–84
88. Mattia M, Del Giudice P (2000) Efficient event-driven simulation of large networks of spiking neurons and dynamical synapses. *Neural Comput* 12:2305–2329
89. Mazor O, Laurent G (2005) Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron* 48:661–673
90. McNaughton B, Battaglia F, Jensen O, Moser E, Moser M (2006) Path integration and the neural basis of the 'cognitive map'. *Nat Rev Neurosci* 7:663–678
91. Mongillo G, Curti E, Romani S, Amit DJ (2005) Learning in realistic networks of spiking neurons and spike-driven plastic synapses. *Eur J Neurosci* 21:3143–3160
92. Moreno-Bote R, Rinzel J, Rubin N (2007) Noise-induced alternations in an attractor network model of perceptual bistability. *J Neurophysiol* 98:1125–1139
93. Murphy TH, Blatter LA, Wier WG, Baraban JM (1992) Spontaneous synchronous synaptic calcium transients in cultured cortical neurons. *J Neurosci* 12:4834–45
94. O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. *Exp Brain Res* 34:171–175
95. O'Keefe J, Recce M (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* 3:317–330
96. Opitz T, Lima ADD, Voigt T (2002) Spontaneous development of synchronous oscillatory activity during maturation of cortical neurons in vitro. *J Neurophysiol* 88:2196–206
97. Perez-Orive J, Mazor O, Turner G, Cassenaer S, Wilson R, Laurent G (2002) Oscillations and sparsening of odor representations in the mushroom body. *Science* 297:359–365
98. Petersen C, Grinvald A, Sakmann B (2003) Spatiotemporal dynamics of sensory responses in layer 2/3 of rat barrel cortex measured in vivo by voltage-sensitive dye imaging combined with whole-cell voltage recordings and neuron reconstructions. *J Neurosci* 23:1298–1309
99. Pikovsky A, Rosenblum M, Kurth J (2001) Synchronization, a universal concept in nonlinear science. Cambridge University Press, Cambridge
100. Polonsky A, Blake R, Braun J, Heeger DJ (2000) Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nat Neurosci* 3:1153–1159
101. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C. Cambridge University Press, Cambridge
102. Renart A, Moreno R, Wang X-J (2003) Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron* 38:473–485
103. Reyes A, Fetz E (1993) Effects of transient depolarizing potentials on the firing rate of cat neocortical neurons. *J Neurophysiol* 69:1673–1683
104. Richardson MJE, Brunel N, Hakim V (2003) From subthreshold to firing-rate resonance. *J Neurophysiol* 89:2538–2554

105. Risken H (1984) The Fokker–Planck equation: methods of solution and applications. Springer, Berlin
106. Roelfsema P, Lamme V, Spekreijse H (2004) Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nat Neurosci* 7:982–991
107. Sanchez-Vives MV, McCormick DA (2000) Cellular and network mechanisms of rhythmic recurrent activity in neocortex. *Nat Neurosci* 3:1027–34
108. Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J Neurosci* 18:3870–3896
109. Shu Y, Hasenstaub A, McCormick DA (2003) Turning on and off recurrent balanced cortical activity. *Nature* 423:288–93
110. Sjöström PJ, Turrigiano GG, Nelson S (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron* 32:1149–1164
111. Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J Neurosci* 13:334–350
112. Stein R (1965) A theoretical analysis of neuronal variability. *Biophys J* 5:173–194
113. Stopfer M, Jayaraman V, Laurent G (2003) Intensity versus identity coding in an olfactory system. *Neuron* 39:991–1004
114. Tong F, Nakayama K, Vaughan JT, Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate visual cortex. *Neuron* 21:761–773
115. Tsodyks MV, Kenet T, Grinvald A, Arieli A (1999) Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* 286:1943–1946
116. Tsodyks MV, Skaggs W, Sejnowski T, McNaughton B (1996) Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. *Hippocampus* 6:271–280
117. Tsodyks MV, Pawelzik, Markram H (1998) Neural networks with dynamic synapses. *Neural Comp* 10:821–835
118. Tsodyks MV, Uziel A, Markram H (2000) Synchrony generation in recurrent networks with frequency-dependent synapses. *J Neurosci* 20:RC50
119. Turrigiano GG, Nelson SB (2000) Hebb and homeostasis in neuronal plasticity. *Curr Opin Neurobiol* 10:358–364
120. van Vreeswijk C, Sompolinsky H (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274:1724–1726
121. van Vreeswijk C, Sompolinsky H (1998) Chaotic balanced state in a model of cortical circuits. *Neural Computation* 10:1321–1371
122. van Vreeswijk C, Abbott L, Ermentrout GB (1994) When inhibition not excitation synchronizes neural firing. *J Comput Neurosci* 1:313–321
123. Wallén P et al (2007) Sodium-dependent potassium channels of a slack-like subtype contribute to slow afterhyperpolarization in lamprey spinal neurons. *J Physiol* 585:75–90
124. Wilson HR (2003) Computational evidence for a rivalry hierarchy in vision. *Proc Natl Acad Sci USA* 100:14499–14503
125. Wilson M, McNaughton B (1993) Dynamics of the hippocampal ensemble code for space. *Science* 261:1055–1058
126. Wyart C, Cocco S, Bourdieu L, Léger JF, Herr C, Chatenay D (2005) Dynamics of excitatory synaptic components in sustained firing at low rates. *J Neurophysiol* 96:3370–80
127. Zillmer R, Brunel N, Hansel D (2008) Irregular states in randomly diluted networks of leaky integrate-and-fire neurons. in preparation.

## Books and Reviews

- Buzsaki G (2006) Rhythms of the brain. Oxford University Press
- Chow C, Gutkin B, Hansel D, Meunier C, Dalibard J (eds) (2004) Methods and models in neurophysics, Les Houches 2003. North-Holland
- Dayan P, Abbott L (2001) Theoretical neuroscience. MIT Press, Cambridge
- Gerstner W, Kistler WM (2002) Spiking neuron models: single neurons, populations, plasticity. Cambridge University Press, Cambridge
- Wang X-J (2003) Neural oscillations. In: Nadel L (ed) Encyclopedia of Cognitive Science. MacMillan, London, pp 272–280

## Noise and Stability in Modelocked Soliton Lasers

BRIAN H. KOLNER<sup>1,2</sup>

<sup>1</sup> Department of Applied Science,  
University of California, Davis, USA

<sup>2</sup> Department of Electrical and Computer Engineering,  
University of California, Davis, USA

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 List of Symbols  
 Basic Laser Physics  
 Laser Dynamics  
 Laser Instability, Chaos and the Lorenz Equations  
 Modelocking  
 Solitons  
 Modelocked Soliton Lasers  
 Laser Noise and Linewidth  
 Noise in Soliton Lasers  
 Large Scale Instability in Soliton Lasers  
 Laser Clocks  
 Conclusions and Outlook for the Future  
 Acknowledgments  
 Bibliography

## Glossary

**First and second laser thresholds** Minimum conditions for producing stable and chaotic laser operation, respectively.



**Lorenz equations** A system of three coupled nonlinear differential equations describing convective fluid flow in cells. These equations and their solutions launched the field of chaos theory.

**Modelocked laser** A laser with evenly spaced modes which have their phases locked together so that the superposition of the modes creates a periodic train of very short pulses ( $< 1$  ps).

**Poisson photon distribution** Probability of  $k$  photons arriving in a given interval of time for which there are an average of  $\bar{n}$  photons.

**Relaxation oscillation** Periodic fluctuation about an equilibrium point when a system, initially operating in steady-state, is subjected to a transient perturbation. The system “relaxes” back to equilibrium through a (usually) damped sinusoidal response.

**Shot noise** Noise produced by the random arrival of electrons in a photodetector illuminated by a laser field.

**Soliton** A solitary wave packet which is a solution to a nonlinear wave equation in a medium possessing both dispersion and nonlinearity. These effects balance exactly so that the wave packet maintains its shape and is stable against perturbations.

## Definition of the Subject

Since its conception in 1958, the laser has become a nearly ubiquitous feature of the technological landscape finding its way into a spectacular array of devices from data storage to entertainment to fundamental science and metrology and myriad others. One of the principal features of the laser is its remarkable coherence. It is a nearly ideal source of monochromatic radiation and from this property the fields of precision spectroscopy and timekeeping have reaped huge rewards. Recently, an innovation in the design of a class of ultrashort-pulse lasers has resulted in a new type of optical clockwork with stability that rivals and even supercedes that of the best atomic clocks, including the cesium-beam clocks which, by international agreement, are used to define the second. The nearly perfect train of optical pulses emitted from these laser clocks are associated with a nearly ideal spectrum of periodic spikes, or comb-lines, in the optical frequency domain which can act as a “frequency ruler” for precision measurements. In just a few short years this new type of laser has revolutionized the fields of precision spectroscopy and timekeeping and there is much more yet to come. This raises the interesting question “how stable can these new laser clocks really be?”. To begin to address this we must consider the fundamental mechanisms of noise and stability of lasers. This is the subject of the present chapter.

## Introduction

Lasers are wonderfully complex systems incorporating energy storage and conversion, interaction of radiation with ensembles of atoms or molecules, classical optics, feedback and linear and nonlinear processes everywhere. Fortunately, most of the processes involved in laser action are amenable to approximations that make the system analytically tractable, although they can frequently mask very interesting and often technologically important phenomena.

The most striking feature of the laser is the ability to combine the electromagnetic radiation of, typically,  $10^{15}$  atomic oscillators in a coherent fashion and produce a single narrow beam of light at a single frequency. The property of any radiation source that describes its power per-unit-area, per-unit-frequency interval, per-solid-angle is called the *brightness*. No other source of electromagnetic radiation in the visible portion of the spectrum comes close to the brightness of a laser.

The second most striking feature of the laser after its brightness is its degree of spatial and temporal coherence. By this we mean the stability in space and time of the harmonic nature of the electromagnetic field. Coherence is typically characterized by interfering two portions of wavefronts separated either in space or in time. High spatial coherence allows the laser field to be focused to a tight spot while high temporal coherence allows the field to accurately probe very narrow spectroscopic features of atoms or molecules or to serve as a frequency or wavelength standard. Although both of these properties influence the brightness, we will concern ourselves in this chapter exclusively with the temporal properties of the field.

To study the temporal behavior of any oscillator it will be useful to introduce the concepts of amplitude and phase instability. A perfect, unit-amplitude, harmonic oscillator has an output (field, voltage, current, position, etc.) given by the simple function  $E_o(t) = \sin(\omega_o t)$ . No physical oscillator displays perfect amplitude and phase stability so a more realistic function could be written as

$$E_n(t) = (1 + \epsilon(t)) \sin(\omega_o t + \phi(t)) \quad (1)$$

where  $\epsilon(t)$  and  $\phi(t)$  are random processes describing the effects of amplitude and phase noise, respectively. In “good” oscillators, these random processes do not cause dramatic deviations from the ideal behavior and if one measures the waveform in the time domain, they are seldom observable. However, if the waveform is analyzed for its frequency content, even tiny amounts of amplitude or phase noise can be readily measured. In fact, almost all oscillator characterization is done by measuring the frequency spectrum except for the case of the very most sta-



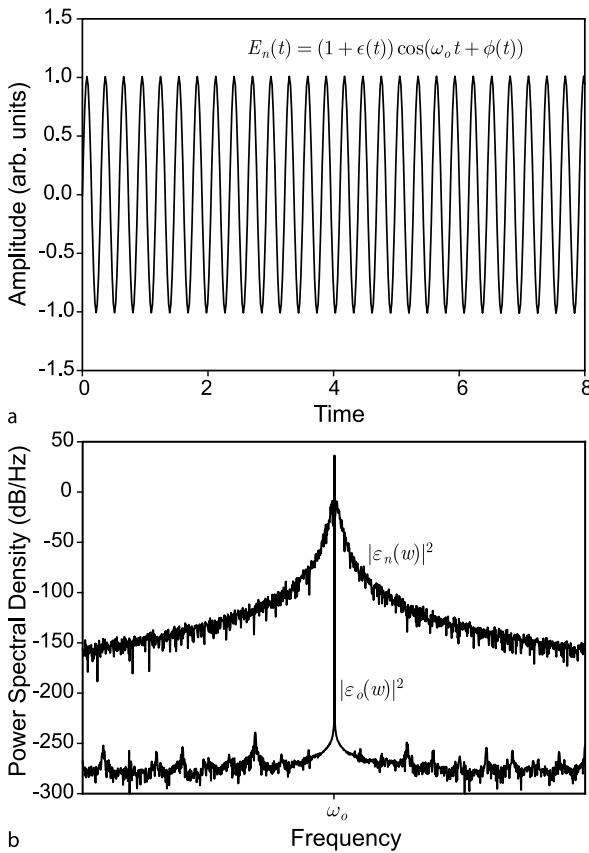
ble sources (e. g. atomic clocks and the new self-referenced laser clocks) where the statistics of the timing fluctuation are of greatest interest and the time domain yields valuable insight. In both cases very long time records are acquired. The domain of analysis is then predicated on whether the noise processes have slowly (time-domain statistics) or rapidly (frequency-domain power spectra) evolving components [1,2,3].

Figure 1a shows a simple sine-wave with a small amount of amplitude and phase noise of the form described by (1). For this example  $\epsilon(t)$  and  $\phi(t)$  are zero-mean Gaussian random processes with standard deviations of 0.2 and  $0.02\pi$ , respectively. In addition, the spectral content of the noise was low-pass filtered to emphasize the low frequency content. Notice that the time-domain representation of the function displays no observable instability while the power spectrum  $|E_n(\omega)|^2$  shown in Fig. 1b contains a very strong signature of the com-

bined amplitude and phase noise. For reference, the spectrum of the pure sinusoid  $|E_o(\omega)|^2$  is also shown. Its near delta-function form is only marred by the noise floor set by the limited precision of the computer. This figure demonstrates several very important ideas. First, even a very small amount of noise added to the amplitude and phase of an otherwise perfect sinusoid can dramatically alter the purity of its spectrum. Second, once the spectrum is corrupted by the noise, precise establishment of the center frequency of the oscillator is now complicated by the spreading of the energy in frequency space. This so-called “linewidth broadening” is present in all oscillators and a thorough study of the causative mechanisms is essential for good oscillator design. Finally, the spreading of energy into noise sidebands adjacent to the main peak can completely mask the presence of other desired signals in that region. For this reason, spectral purity of oscillators is a very important quality in the design and realization of receivers at any frequency from radio through microwave and optical portions of the spectrum.

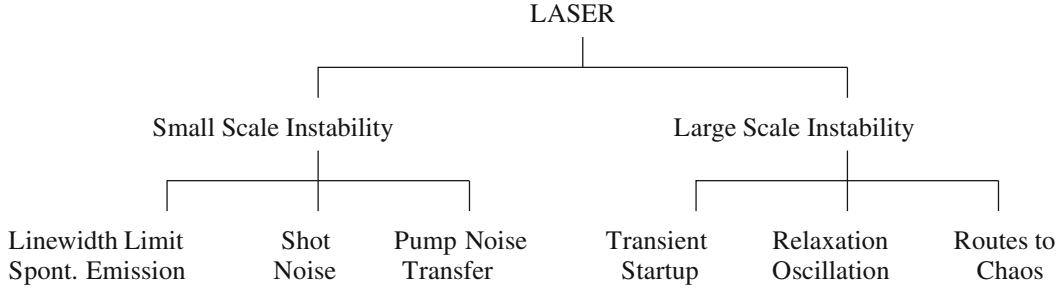
The addition of small amounts of random noise to the amplitude and phase of a laser field is but one kind of instability which we will call “small scale”. The fluctuations are typically very small compared to the average quantities and the laser operates in a steady-state regime with essentially constant power flow into and out of the laser system and constant energy-level populations. Examples of processes that produce small-scale amplitude and phase noise include spontaneous emission of photons into the desired cavity mode in addition to the stimulated emission photons created by lasing action. This gives rise to a fundamental linewidth limit (i. e. the Schawlow–Townes limit). The discrete and random generation of charge carriers in a photon detector gives rise to “shot noise” when detecting laser fields. Also, amplitude noise present on the pump source will be transferred to the laser creating both amplitude and phase noise in the final output field. These are summarized in Fig. 2.

There is another type of instability which we will call “large-scale” and this is characterized by large, and sometimes chaotic, swings in the principal laser parameters such as electric field, population inversion and atomic polarization. The first instance of these phenomena always occurs when a laser first starts up as the population inversion approaches threshold for oscillation. The closely related phenomenon of relaxation oscillation occurs when a sudden change in the pumping rate or intracavity loss occurs and upsets the steady-state balance between pumping, stimulated emission and output coupling. Finally, in certain regions of parameter space, the laser can be operated in a wildly unstable manner that is quite analogous



**Noise and Stability in Modelocked Soliton Lasers, Figure 1**

**a** Sinewave oscillator output with small amount of amplitude and phase noise. **b** Power spectrum of oscillator output  $|E_n(\omega)|^2$ . Power spectrum of a pure sinusoid,  $|E_o(\omega)|^2$ , (i. e.  $\epsilon(t) = \phi(t) = 0$ ) shown for comparison



Noise and Stability in Modelocked Soliton Lasers, Figure 2

Categorization of noise and instabilities associated with laser oscillators

to nonlinear dissipative systems undergoing chaotic dynamics. All of the phenomena mentioned and shown in Fig. 2 will be discussed in this chapter. Many of the topics can be found in textbooks [4,5,6,7,8,9,10,11] and comprehensive monographs [12,13,14,15,16,17,18,19] and some, which will be presented toward the end of the chapter, are quite new.

### List of Symbols

$b$  = radius of laser rod (m).  
 $H_{AM}(\omega_m)$  = complex AM-to-AM noise transfer function.  
 $H_{PM}(\omega_m)$  = complex AM-to-PM noise transfer function.  
 $k$  = thermal diffusivity ( $\text{m}^2/\text{s}$ ).  
 $K$  = photon number/population inversion density coupling coefficient ( $\text{m}^3 \text{s}^{-1}$ ).  
 $L_c$  = cavity length without gain medium (m).  
 $L_a$  = length of gain medium (m).  
 $n(t)$  = cavity photon number density ( $\text{m}^{-3}$ ).  
 $\tilde{n}_1$  = complex photon number density phasor amplitude ( $\text{m}^{-3}$ ).  
 $N(t)$  = population inversion density ( $\text{m}^{-3}$ ).  
 $\tilde{N}_1$  = complex population inversion density phasor amplitude ( $\text{m}^{-3}$ ).  
 $p_c$  = round trip cavity distance (m).  
 $p_m$  = round trip distance through gain medium (m).  
 $P_{L0}$  = steady-state laser output power (W).  
 $P_{p0}$  = steady-state pump power (W).  
 $r$  = normalized pumping rate above threshold.  
 $r'$  = radial coordinate of laser rod (m).  
 $R_p(t)$  = effective pumping rate density into upper laser level ( $\text{m}^{-3} \text{s}^{-1}$ ).  
 $R_{p0}$  = steady-state pumping rate density into upper laser level ( $\text{m}^{-3} \text{s}^{-1}$ ).  
 $R_{p1}$  = peak sinusoidal amplitude of pumping rate density ( $\text{m}^{-3} \text{s}^{-1}$ ).  
 $R_{\text{tot}}$  = product of mirror reflectivities.  
 $T$  = round trip cavity time (s).  
 $T_{\text{oc}}$  = power transmission of output coupler.

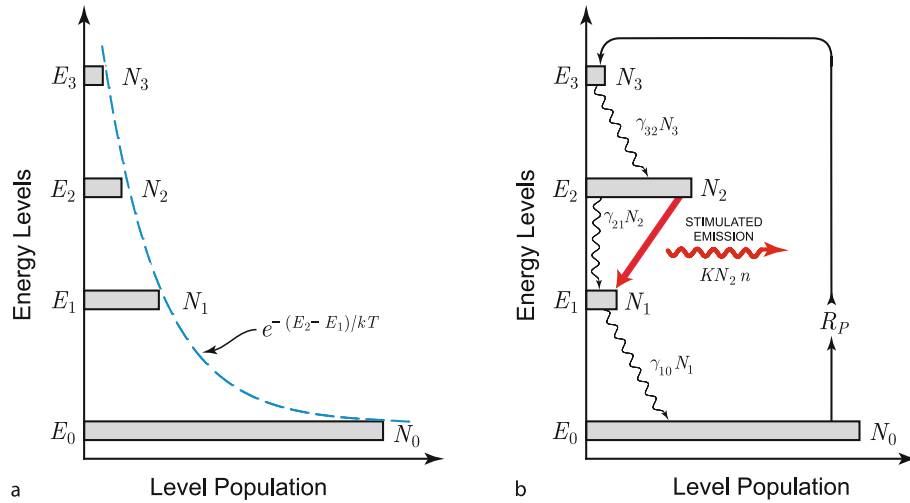
$V_c$  = cavity mode volume including gain medium ( $\text{m}^3$ ).  
 $V_g$  = mode volume of gain medium ( $\text{m}^3$ ).  
 $\alpha_0$  = ohmic loss coefficient ( $\text{m}^{-1}$ ).  
 $\beta(\omega_m)$  = peak phase deviation.  
 $\beta_1$  = eigenvalues of the radial heat diffusion problem.  
 $\Delta P_L$  = peak fractional laser power deviation.  
 $\Delta P_p$  = peak fractional pump power deviation.  
 $\Delta\omega_a$  = gain bandwidth, FWHM ( $\text{s}^{-1}$ ).  
 $\gamma_c$  = cavity decay rate =  $\frac{2\alpha_0 p_c + \ln(1/R_{\text{tot}})}{T}$  ( $\text{s}^{-1}$ ).  
 $\gamma_{\text{rad}}$  = radiative decay rate out of upper level ( $\text{s}^{-1}$ ).  
 $\gamma_2$  = total upper level decay rate ( $\text{s}^{-1}$ ).  
 $\eta$  = laser slope efficiency.  
 $\omega_a$  = resonant frequency of active gain medium ( $\text{s}^{-1}$ ).  
 $\omega_m$  = pump modulation frequency ( $\text{s}^{-1}$ ).  
 $\tau_{\text{co}}$  = unperturbed round-trip cavity time.  
 $\tau_p$  = modelocked pulsewidth.  
 $3^*$  = polarization-field alignment factor.

### Basic Laser Physics

Lasers represent a very complex system of interrelated objects and phenomena. In order to understand the origin of the systems of equations that can predict laser behavior, we should first review the basic principles of laser operation.

The fundamental mechanism of lasers is the stimulated emission of electromagnetic waves from atoms or molecules which are temporarily in an excited state. That is, a configuration corresponding to an energy level that is higher than the atom would normally find itself in thermal equilibrium. From a statistical standpoint, when a large number of atoms are gathered together and share a common temperature  $T$ , the probability of an atom being in any energy level is given by the Boltzmann distribution. This tells us that the ratio of the number of atoms,  $N_2$ , in energy level  $E_2$ , to the number of atoms,  $N_1$  in a lower energy level,  $E_1$  is given by

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{kT}\right) = \exp\left(-\frac{\hbar\omega_a}{kT}\right) \quad (2)$$



**Noise and Stability in Modelocked Soliton Lasers, Figure 3**

Energy levels  $E_i$  and corresponding populations  $N_i$  for the case of a thermal equilibrium with no pumping or driving signals and **b** pumping with rate  $R_P$  from the lowest energy level  $E_0$  through  $E_3$  to the upper laser level  $E_2$ . Relaxation rates  $\gamma_{ij}N_i$  and stimulated emission  $KN_2n$  compete with the pumping rate to form a closed system described by rate equations

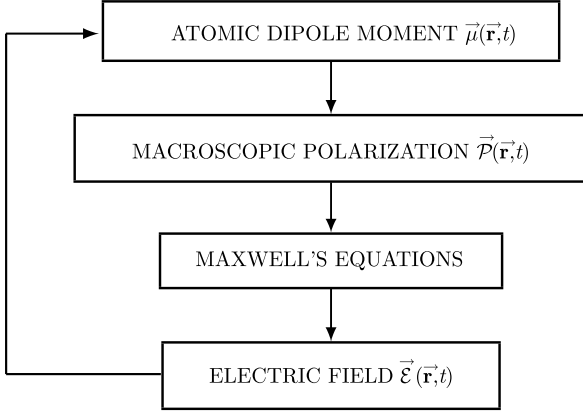
where  $k = 1.381 \times 10^{-23} \text{ J K}^{-1}$  is Boltzmann's constant,  $\hbar = 6.626 \times 10^{-34} \text{ Js}$  is Planck's constant and  $\omega_a \equiv (E_2 - E_1)/\hbar$  is the radian frequency of electromagnetic radiation corresponding to the transition between the energy levels. Any pair of energy levels must have this same exponential relationship between the corresponding populations as shown in Fig. 3a.

In order to change a medium from absorbing radiation between two energy levels to contributing energy to a radiation field (and thus exhibit amplification), there must be more atoms in a higher energy level, or state, than in a lower level. This, of course, no longer constitutes a situation of thermal equilibrium and must be created by an outside agent. The preparation of an excited state is brought about by pumping the atoms from a lower energy state to a higher energy state through some energy transfer mechanism. Figure 3b shows how energy is transferred from the ground state  $E_0$  to the higher energy state  $E_3$  at a rate  $R_P$ . This diagram depicts a so-called four-level laser where the uppermost level exhibits rapid relaxation,  $\gamma_{32}N_3$ , to the upper laser level  $E_2$ , where the atoms stay for a considerable time. If the relaxation from the lower laser level  $\gamma_{10}N_1$  is also much faster than relaxation from  $E_2 \rightarrow E_1$ , then the population  $N_2$  can be made to exceed  $N_1$  creating a "population inversion". Under these circumstances, an electromagnetic wave can experience amplification by the process of stimulated emission. Stimulated emission of a quantum of electromagnetic energy occurs when the atom is driven by an electromagnetic field whose frequency is matched to the energy difference between two

levels for which a population inversion exists. The remarkable feature of the stimulated emission is that it is coherent with the driving field. This is also an essential feature of laser action since any oscillator requires both gain and feedback of coherently superposed waves. The feedback, in the case of a laser, is provided by mirrors which return most of the amplified fields back to the atoms to sustain the stimulated emission process in concert with the pumping action.

As we shall see, the dynamics of laser operation are controlled by the time constants of many competing processes. Some of these processes are microscopic and fundamental (e.g. atomic lifetimes), and many are macroscopic and environmental (e.g. Doppler effects, crystal strain broadening, cavity geometry, etc.). The multitude of different time constants, energy supply and depletion rates and the (usually) very large number of atoms required to deliver useful power output suggests that there is great potential for cooperative effects between the ensemble of atomic oscillators and the electromagnetic fields which drive them and to which they contribute.

Although the coupling of radiation to atoms is most rigorously described using quantum mechanics, when large numbers of atoms are involved a "mostly classical" treatment is satisfactory and brings out almost all of the important phenomena of laser dynamics without requiring the extra machinery of quantum physics. The one concession we make to the quantum world is reliance on the notion of defined energy levels or eigenstates of the atoms. It is the transitions between these states that characterize



**Noise and Stability in Modelocked Soliton Lasers, Figure 4**  
Self-consistent flow model describing coupling of atomic oscillators to electromagnetic fields

the frequencies of radiation, either absorbed or emitted, from the oscillator.

In order to gain a little insight into the physics of laser action, we can step back and take a broader view. The simplified flow model shown in Fig. 4 describes a self-consistent picture of the coupled processes that combine the radiation from individual atomic or molecular oscillators. In the classical picture of the interaction of radiation with matter, the electromagnetic wave causes a small time-varying displacement of the atom's charge cloud. The product of this displacement and the net charge displaced is called the *dipole moment*;  $\vec{\mu}(\vec{r}, t) = -e\vec{x}(\vec{r}, t)$  where  $e$  is the charge on a single electron and  $\vec{x}(\vec{r}, t)$  is the separation between the charge cloud and its positively charged ion located at position  $\vec{r}$ . These microscopic dipole moments are akin to tiny bar magnets (except that these are electric dipole fields) and thus a large number of them in a macroscopic volume can have a significant effect on the local total electric field. An ensemble of microscopic dipole moments collectively makes up the polarization field  $\vec{P}(\vec{r}, t)$ , which is the sum of all the individual dipole moments divided by the volume enclosing those dipoles (therefore a "dipole moment per-unit-volume"). The polarization field combines with the electric field to yield the so-called displacement field  $\vec{D}(\vec{r}, t)$  which is part of Maxwell's famous equations of electromagnetism [20]. In fact, one can show that the polarization field acts as a driving term for the differential wave equation for the electric field. The solution to the wave equation is the electric field that is responsible for establishing the individual dipole moments in the first place and we are back at the beginning of our flow chart.

Now we begin a little more detailed analysis to discover the basic dynamics of the material polarization which is

responsible for laser action. The principles that are developed in the following section can be found in any good textbook on optics or laser physics but often the notation is inconsistent, which can cause confusion. Here we mostly follow the approach and notation of Siegman [4].

The equation of motion for the dipole moment  $\vec{\mu}(\vec{r}, t)$  due to a driving field  $\vec{E}(\vec{r}, t)$  is that of the well-known damped harmonic oscillator. In the context of an electron bound to a positively-charged nucleus, this is frequently referred to as the "classical electron oscillator" or CEO model;

$$\frac{d^2\mu_x(t)}{dt^2} + \gamma \frac{d\mu_x(t)}{dt} + \omega_a^2\mu_x(t) = \frac{e^2}{m}\mathcal{E}_x(t) \quad (3)$$

where  $\gamma$  is the damping term or energy decay rate (about which more will be said shortly),  $\omega_a$  is the atomic resonant frequency, and, for simplicity, we have written just one of the scalar components of the full vector system. In the absence of a forcing function,  $\mathcal{E}_x(t) = 0$ , the natural response for the single dipole is

$$\mu(t) = \mu_o \exp\left[-\frac{\gamma}{2}t + i\omega'_a t\right] \quad (4)$$

where  $\omega'_a$  is the actual oscillation frequency in the presence of damping

$$\omega'_a \equiv \sqrt{\omega_a^2 - (\gamma/2)^2} \quad (5)$$

and  $\omega_a$  is the natural oscillation frequency of the atom in the CEO model. In the quantum picture  $E_{21} = \hbar\omega_a$  is the energy difference between the excited and ground state energy levels. If we calculate the purely radiative decay rate for a classical electron oscillator

$$\gamma_{\text{rad,ceo}} = \frac{e^2\omega_a^2}{6\pi\epsilon_0 mc^3} \quad (6)$$

we find a typical value of  $\gamma_{\text{rad,ceo}} \approx 10^8 \text{ s}^{-1}$  [4]. Compared to the optical resonant frequency of  $\omega_a \approx 2\pi \times 10^{15} \text{ s}^{-1}$  it is reasonable to set  $\omega'_a = \omega_a$  for all future discussions.

When we add up the dipole moments of an ensemble of atoms with density  $N$  in a small volume  $V$ , we obtain the dipole moment-per-unit volume, or, the *polarization*  $\vec{P}(\vec{r}, t) = \sum_{i=1}^{NV} \vec{\mu}_i(\vec{r}, t)$ . Like the electric field, the polarization is a vector field with space and time dependence. The importance of the material polarization becomes clear when we realize that once the dipoles are set oscillating by an incident field, they become source terms for additional radiation and thus modify the total electric field. This can be summarized elegantly in the wave equation for the electric field which is derived from Maxwell's equations

$$\nabla^2 \vec{E}(\vec{r}, t) - \mu_o \epsilon_o \frac{\partial^2 \vec{E}(\vec{r}, t)}{\partial t^2} = \mu_o \frac{\partial^2 \vec{P}(\vec{r}, t)}{\partial t^2}. \quad (7)$$

The term on the right-hand side is the source term or forcing function for the wave equation.

### Natural Lifetime and Dephasing

An important attribute of both the dipole moment and polarization is the exponential decay time associated with the natural lifetime of a single oscillator and the dephasing of an ensemble of oscillators. The CEO model of the atom includes a phenomenological damping term  $\gamma$  which accounts for energy lost by radiation  $\gamma_{\text{rad}}$  (i. e. spontaneous emission) and nonradiative effects  $\gamma_{\text{nr}}$  such as collisions with other atoms and vessel walls. The two combine to form a net energy decay rate  $\gamma$  such that the internal energy of the atom's electron cloud motion evolves according to

$$U(t) = U_0 \exp(-\gamma t) \quad \text{where} \quad \gamma = \gamma_{\text{rad}} + \gamma_{\text{nr}}. \quad (8)$$

The energy decay rate  $\gamma$  is sometimes seen in the literature as  $\gamma_{\parallel}$  and called the “longitudinal relaxation rate” [6,16] a term held over from the Bloch equations of nuclear magnetic resonance [21].

When an ensemble of simple classical electron oscillators, each with a natural response (4), radiates together we must carefully consider their superposition. If they all start oscillating in phase and there is no disruption, then it is easy to see that the macroscopic polarization field will have the same exponential decay as a single oscillator and is given by the sum

$$\mathcal{P}_x(t) = N\mu_{x0} \exp\left[-\frac{\gamma}{2}t + i\omega_a t\right]. \quad (9)$$

For simplicity we have dropped the vector notation for the polarization field and just consider one Cartesian coordinate.

Unfortunately, real media do not exhibit a continuous fully coherent response as in (9) but instead undergo *dephasing events* such as collisions which tend to randomize the phases of the oscillators with respect to each other. This effect can be accounted for by introducing an additional decay rate term,  $2/T_2$ , which describes the dephasing time of the coherent ensemble, into the equation of motion for the polarization

$$\frac{d^2 \mathcal{P}_x(t)}{dt^2} + \left(\gamma + \frac{2}{T_2}\right) \frac{d\mathcal{P}_x(t)}{dt} + \omega_a^2 \mathcal{P}_x(t) = \frac{Ne^2}{m} \mathcal{E}_x(t). \quad (10)$$

The factor of 2 in  $2/T_2$  comes about because the phenomenon of dephasing is a field effect, not an energy decay. The natural response is

$$\mathcal{P}_x(t) = \mathcal{P}_{x0} \exp\left[-\left(\frac{\gamma}{2} + \frac{1}{T_2}\right)t + i\omega_a t\right]. \quad (11)$$

The total decay rate for the natural response polarization field is now  $\gamma/2 + 1/T_2$ . Equivalently, the  $1/e$  time constant for the decay of the polarization field in  $(\gamma/2 + 1/T_2)^{-1}$ . The time constant  $T_2$  has the very real physical interpretation as the average time an atom oscillates between suffering a dephasing event. The total decay rate (which we will see represents a bandwidth) or more properly, it's inverse, represents the first major time constant responsible for interesting laser dynamics. It is, however, generally dominated by the dephasing rate ( $T_2 \ll 2/\gamma$ ). We note, also, that in the literature of nuclear magnetic resonance, the dephasing rate is often seen as  $\gamma_{\perp}$  and called the “transverse” or polarization relaxation rate [6,16].

### Steady-State Response and the Susceptibility

The characterization of any linear system by differential equations allows one to evaluate a natural, or homogeneous, response (11) and a forced, or inhomogeneous response. A particularly useful forcing function is the steady harmonic source  $e^{i\omega t}$  since the system will respond in kind with a harmonic signal whose amplitude and phase depend on the amplitude and frequency of the driving function. To this end, we introduce the following forcing and response functions

$$\mathcal{E}_x(t) = \text{Re}[\tilde{E}_x e^{i\omega t}] = \frac{1}{2} [\tilde{E}_x e^{i\omega t} + \tilde{E}_x^* e^{-i\omega t}] \quad (12)$$

$$\mathcal{P}_x(t) = \text{Re}[\tilde{P}_x e^{i\omega t}] = \frac{1}{2} [\tilde{P}_x e^{i\omega t} + \tilde{P}_x^* e^{-i\omega t}] \quad (13)$$

into Eq. (10). The solution for the polarization amplitude is therefore dependent on the driving frequency according to

$$\tilde{P}(\omega) = \frac{Ne^2}{m} \frac{1}{\omega_a^2 - \omega^2 + i\omega(\gamma + 2/T_2)} \tilde{E}(\omega). \quad (14)$$

This polarization field now becomes a driving source term in Maxwell's equations through the displacement field

$$\tilde{D}(\omega) = \epsilon_0 \tilde{E}(\omega) + \tilde{P}(\omega). \quad (15)$$

The atoms involved in laser action usually make up only a small fraction of the total matter density of the active medium. The remaining host atoms also have atomic resonances and contribute a polarization response but the resonant frequencies are usually far removed from the laser transitions and the contribution to the polarization is therefore rather constant throughout the range of the laser atom's response. Therefore, it is useful to separate



out the polarization due to the host medium from the laser atoms's atomic response and write

$$\tilde{P}(\omega) = \tilde{P}_{\text{host}}(\omega) + \tilde{P}_{\text{at}}(\omega) \quad (16)$$

which then results in a modified constitutive relation

$$\tilde{D}(\omega) = \epsilon_0 \tilde{E}(\omega) + \tilde{P}_{\text{host}}(\omega) + \tilde{P}_{\text{at}}(\omega) = \epsilon_{\text{host}} \tilde{E}(\omega) + \tilde{P}_{\text{at}}(\omega). \quad (17)$$

Since the atomic polarization depends linearly on the driving field, we can define a scale factor which we call the "complex atomic susceptibility" where

$$\tilde{\chi}(\omega) = \frac{\tilde{P}_{\text{at}}(\omega)}{\epsilon_{\text{host}} \tilde{E}(\omega)} = \frac{Ne^2}{\epsilon_{\text{host}} m} \frac{1}{[\omega_a^2 - \omega^2 + i\omega(\gamma + 2/T_2)]}. \quad (18)$$

The frequency dependence of  $\tilde{\chi}(\omega)$  represents the classic "Lorentzian lineshape" function, the real part of which describes the wave speed while the imaginary part the loss or gain of the medium. The characteristic width of this lineshape is typically very small compared to the atomic transition frequency which allows us to use the resonance approximation and write (18) as

$$\tilde{\chi}(\omega) = \chi' + i\chi'' \approx -\chi_0'' \left[ \frac{\Delta x}{1 + \Delta x^2} + i \frac{1}{1 + \Delta x^2} \right] \quad (19)$$

where

$$\chi_0'' \equiv \frac{Ne^2}{m\omega_a \epsilon \Delta\omega_a}, \quad \Delta x \equiv 2 \frac{\omega - \omega_a}{\Delta\omega_a} \quad (20)$$

is the normalized detuning from line center and

$$\Delta\omega_a \equiv \gamma + 2/T_2 \quad (21)$$

is the linewidth measured at the full-width, half maximum point (FWHM). The two components of the linewidth are the energy decay rate,  $\gamma$ , and the inverse of the dephasing time  $T_2$ . In most laser systems the linewidth is dominated by dephasing.

When the polarization field (16) is inserted into the wave equation (7) the solution is a monochromatic wave that propagates according to

$$\mathcal{E}(z, t) = \mathcal{E}_0 e^{i\omega t - \Gamma(\omega)z} \quad (22)$$

in one dimension, where

$$\Gamma(\omega) \equiv i\beta \sqrt{1 + \tilde{\chi}_{\text{at}} - i\sigma/\omega\epsilon_{\text{host}}} \quad \text{and} \quad \beta \equiv \omega \sqrt{\mu_0 \epsilon_{\text{host}}}. \quad (23)$$

Now it is straightforward to see how the complex susceptibility influences wave propagation. It is often convenient to simplify the above expression by first noting that in most materials,  $|\tilde{\chi}_{\text{at}}|$  and  $\sigma$  are both very small compared to unity and thus we can expand the square root to first order. Also, we can group terms according to whether they are real or imaginary and label them in the usual manner of wave problems where  $\alpha$  is used to denote gain or loss and  $\beta$  is used to denote phase shift. Thus,

$$\Gamma(\omega) = i(\beta + \Delta\beta_m(\omega)) - \alpha_m(\omega) + \alpha_0 \quad (24)$$

where

$$\Delta\beta_m(\omega) = \beta\chi'(\omega)/2 \quad \text{atomic contribution to phase shift} \quad (25)$$

$$\alpha_m(\omega) = \beta\chi''(\omega)/2 \quad \text{atomic contribution to gain or loss} \quad (26)$$

$$\alpha_0 = \sigma/2\epsilon_{\text{host}}c \quad \text{ohmic losses}. \quad (27)$$

Substituting these into the basic wave propagation solution (22) we have

$$\mathcal{E}(z, t) = \mathcal{E}_0 \exp\left([\alpha_m(\omega) - \alpha_0]z + i\omega t - i[\beta + \Delta\beta_m(\omega)]z\right). \quad (28)$$

The net gain per-unit-distance is given by the real term in the exponent,  $\alpha_m(\omega) - \alpha_0$ . The phase shift per-unit-distance gives us the phase velocity,  $v_{\text{ph}}$ , in the medium through the usual ratio

$$\frac{\omega}{\beta + \Delta\beta_m(\omega)} = \frac{\omega}{\beta(1 + \chi'(\omega)/2)} = \frac{1}{\sqrt{\mu_0 \epsilon_{\text{host}}}(1 + \chi'(\omega)/2)} = v_{\text{ph}}. \quad (29)$$

This expression can be easily separated into the free space velocity of light,  $c$ , the effects of the host medium and the effects of the laser atoms;

$$v_{\text{ph}} = \frac{1}{\sqrt{\mu_0 \epsilon_0} \frac{\epsilon_{\text{host}}}{\epsilon_0} \left(1 + \frac{\chi'(\omega)}{2}\right)} = \frac{c}{\sqrt{\frac{\epsilon_{\text{host}}}{\epsilon_0} \left(1 + \frac{\chi'(\omega)}{2}\right)}} = \frac{c}{n(\omega)}. \quad (30)$$

The index of refraction is thus

$$n(\omega) = \sqrt{\frac{\epsilon_{\text{host}}}{\epsilon_0} \left(1 + \frac{\chi'(\omega)}{2}\right)}. \quad (31)$$

Everything derived to this point has been from a strictly classical point of view and does not yet admit the possibility of gain. To correct this we simply make a few important substitutions [4]. First, comparing the expressions for the midband susceptibility,  $\chi_0''$  (20), and the radiative decay rate for the classical electron oscillator,  $\gamma_{\text{rad,ceo}}$  (6), we can write

$$\chi_0'' = \frac{3}{4\pi^2} \frac{N\lambda^3 \gamma_{\text{rad,ceo}}}{\Delta\omega_a}. \quad (32)$$

The advantage of this form is that it uses only physically measurable quantities. Next, and most important, is to convert this classical expression to a quantum-mechanically correct expression by substituting for the oscillator density  $N$ , the *population difference density*  $\Delta N \equiv N_1 - N_2$ , where  $N_1$  is the population density of the lower energy state and  $N_2$  is the population density of the upper energy state (see Fig. 3). Therefore, (32) becomes

$$\chi_0'' = \frac{3}{4\pi^2} \frac{\Delta N \lambda^3 \gamma_{\text{rad,ceo}}}{\Delta\omega_a}. \quad (33)$$

This form of the on-resonance susceptibility demonstrates how gain enters the classical picture of the interaction of radiation with matter. When a mechanism produces an inverted population, ( $N_2 > N_1$ ), the sign of  $\chi_0''$  changes and thus so does the imaginary part of the index of refraction (31).

### Feedback and Cavity Modes

Any oscillator, be it mechanical, electrical, optical, etc., relies on two conditions to operate; feedback and gain. For a laser, the feedback is accomplished with mirrors, in between which is placed the gain medium. The mirrors constitute an electromagnetic resonator for the optical fields and only certain frequencies, called *eigenmodes*, can exist stably within this resonator. For general resonators enclosing electromagnetic fields, the exact frequencies of the eigenmodes depend on the wavelength of the oscillating field and the geometry and dimensions of the enclosure.

For a laser, in which the sides of the enclosure are open, the exact frequency at which oscillation occurs is determined by satisfying the condition that one round trip of the electromagnetic field inside the resonator reproduces itself exactly, both in phase and amplitude. This requires that the round trip gain exactly cancels the losses and that there are an integer number,  $n$ , of cycles of the field encompassing one round trip. Equivalently, the round-trip phase shift is exactly  $n2\pi$ . Typically, the length of the cavity is such that  $n$  is very large at optical frequencies, of the order  $10^6$ . Every integer  $n$  that satisfies the  $n2\pi$  roundtrip

phase condition results in what is called an “axial” or “longitudinal” mode. The frequency spacing between the axial modes is determined by adding (or subtracting) one more cycle of phase to the round trip so that  $n \rightarrow n \pm 1$ . It is easy to show that, for a resonator with two mirrors separated by a distance  $L$ , the axial mode spacing is given by

$$\Delta\nu_{\text{ax}} = \frac{c}{2L}. \quad (34)$$

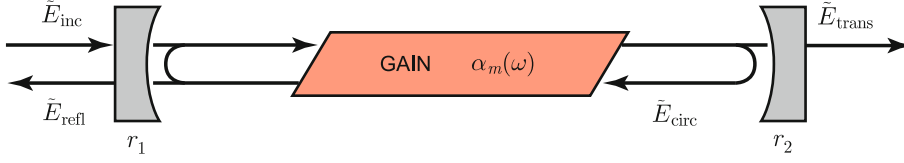
For example, a 1.5 meter-long cavity has an axial mode spacing of 100 MHz. Now, in most cases in the visible and near-infrared portion of the spectrum, the axial mode spacing is smaller than the gain bandwidth of the lasing medium which means that for inhomogeneous media, several modes may oscillate at once. This is typically undesirable since the instantaneous power in the beam fluctuates rapidly due to the beating of the modes. In homogeneously broadened lasers, usually only one mode can oscillate since gain saturation pulls down the gain of all of the atoms uniformly. However, modes may compete for oscillation and this “mode competition” will appear as a fluctuation or instability in the laser output power as different axial modes within the gain curve rapidly rise and fall in power. Many techniques have evolved over the years to force lasers to operate in a single axial mode such as using unidirectional ring cavities, multiple intra-cavity etalons, etc. [9].

All resonators store energy in the cycling fields or mechanical vibrations which they contain. Because of unavoidable loss mechanisms and the need for taking energy out of the resonator, the fields in the resonator have a finite lifetime. The lifetime of the stored energy compared to the period of oscillation defines the quality factor or  $Q$  of the resonator;

$$Q \equiv 2\pi \times \frac{\text{ENERGY STORED}}{\text{ENERGY DISSIPATED PER CYCLE}}. \quad (35)$$

There is an intimate relationship between the  $Q$  of a resonator and the linewidth of an oscillator utilizing the resonator. What we will see is that the gain element in an oscillator must exactly balance any losses in the resonator so that the  $Q$  of the combined resonator plus gain will tend toward infinity. Evidently, this suggests an infinitely narrow linewidth, but, in reality this is prevented by thermal noise in electrical oscillators and quantum noise (spontaneous emission) in optical oscillators.

To gain an appreciation for the elementary properties of a laser resonator and how it contributes to the linewidth-limiting nature of an oscillator below threshold, consider the simple two-mirror laser cavity with gain region shown in Fig. 5. This configuration is called a *regenerative amplifier*. A signal input to the cavity from the



**Noise and Stability in Modelocked Soliton Lasers, Figure 5**  
Schematic diagram of a regenerative laser amplifier

left hand side,  $E_{\text{inc}}$ , passes through a partially transmitting mirror with amplitude reflectivity  $r_1$ . Mirror reflectivity is defined in terms of a field  $E_i$  incident upon the mirror and the subsequent field  $E_r$  reflected by the mirror,  $E_r = r * E_i$ . The mirror amplitude transmission,  $t$ , is defined in terms of the transmitted field  $E_t$  and an incident field according to  $E_t = it * E_i$  [4]. For lossless mirrors, power conservation requires  $r^2 + t^2 = 1$  and we define the power transmission and reflection coefficients as  $T = t^2$  and  $R = r^2$ , respectively.

The wave experiences amplification and phase shift in the gain medium along with losses due to scattering, absorption and transmission through the mirrors. We can follow the wave around one complete round trip of the cavity (length =  $p$ ) including a round trip through the gain (length =  $p_m$ ) and find that

$$E_1 = E_0 \cdot r_1 r_2 \cdot \exp[\alpha_m(\omega)p_m - \alpha_0 p - i\omega p/c - i\Delta\beta_m(\omega)p_m] \quad (36)$$

where  $E_0$  is the field just after passing through the first mirror and  $E_1$  is the field after one round trip.  $\alpha_m(\omega)$  is the frequency-dependent atomic gain (26),  $\alpha_0$  represents the losses due to scattering and background absorption (27) and  $\Delta\beta_m$  is the phase shift due to the atomic medium (25).

From (36) we can define a complex round trip gain

$$\tilde{g}_{\text{rt}}(\omega) = \frac{E_1}{E_0} = r_1 r_2 \cdot \exp[\alpha_m(\omega)p_m - \alpha_0 p - i\omega p/c - i\Delta\beta_m(\omega)p_m] \quad (37)$$

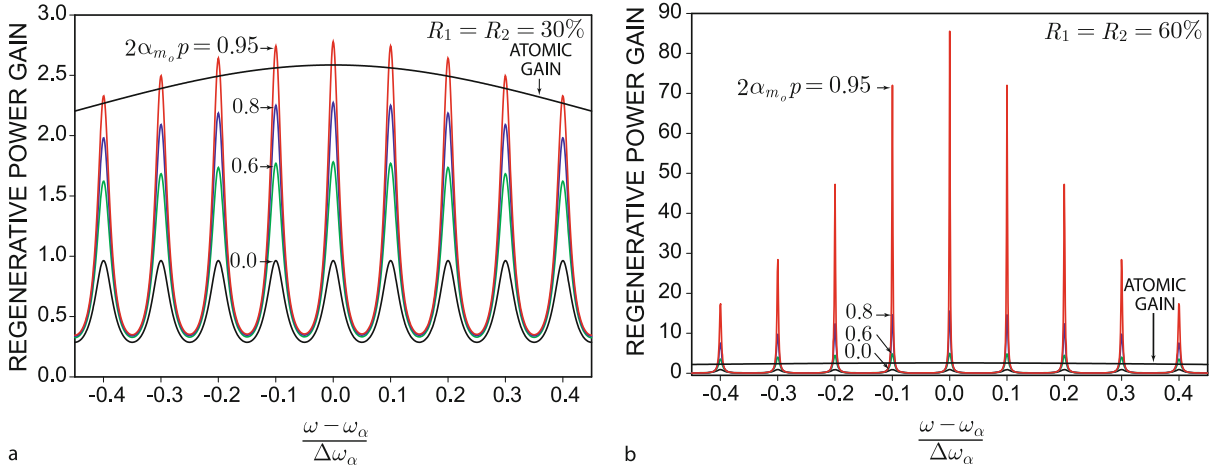
At the second mirror, some of the field leaks out of the cavity and we can consider the overall or net gain of the system as  $\tilde{E}_{\text{trans}}/\tilde{E}_{\text{inc}}$ . However, since some of the signal is reflected back into the cavity and gain medium, a circulating field is built up which can (and usually does) increase far beyond the amplitude of the incident field. This is the notion of *regenerative gain* and the remarkable phenomenon of line-narrowing occurs in this scenario. It is easy to understand qualitatively if we re-examine the ex-

pression for  $Q$  (35). The losses within the cavity, including mirror transmission, degrade the  $Q$  and therefore result in a broad resonance curve or transmission peak. As the atomic gain,  $\alpha_m(\omega)$ , is turned up it begins to compensate for the losses and this improves the  $Q$  and narrows the lineshape or transmission resonance of the composite system. If we follow a wave around the cavity and demand self-consistent fields at all points (i. e. steady-state conditions), we find that the overall amplitude gain of the regenerative amplifier is given by

$$\frac{\tilde{E}_{\text{trans}}}{\tilde{E}_{\text{inc}}} = -\frac{t_1 t_2}{\sqrt{r_1 r_2}} \left( \frac{\sqrt{\tilde{g}_{\text{rt}}(\omega)}}{1 - \tilde{g}_{\text{rt}}(\omega)} \right) \quad (38)$$

Both the overall gain and the line-narrowing really occur due to the denominator in (38). As the round-trip gain approaches unity at a cavity resonance ( $2\pi n$  round-trip phase condition), the denominator is quickly driven toward zero and thus the gain increases dramatically. The closer the magnitude of  $\tilde{g}_{\text{rt}}(\omega)$  is to unity the more quickly this happens as we tune to a cavity resonance, thus producing a narrower response. We can see this effect in Fig. 6 where the overall regenerative power gain,  $|\tilde{E}_{\text{trans}}/\tilde{E}_{\text{inc}}|^2$ , is plotted as a function of frequency for two cases of mirror reflectivities and different values of midband atomic gain  $2\alpha_{m_0}p_m$ . Details of the parameters and quantitative features of the curves are presented in Table 1.

Figure 6a shows the regenerative power gain for mirror reflectivities  $R_1 = R_2 = 30\%$ . The axial mode spacing is chosen, arbitrarily, to be 1/10th of the Lorentzian linewidth  $\Delta\omega_a$ . Notice the qualitative changes in the transmission peaks as the midband atomic gain is varied from  $2\alpha_{m_0}p_m = 0$  to  $2\alpha_{m_0}p_m = 0.95$ . There is a clear trend toward line narrowing. The regenerative amplifier results in an amplifying filter which suppresses noise or unwanted radiation except at the axial mode frequencies. As the gain is increased, even by a small amount, the whole system becomes ever more sensitive to detuning from the axial mode peaks. If we change the mirrors from a reflectivity of  $R = 30\%$  to  $R = 60\%$ , a dramatic line-narrowing effect takes place, as shown in Fig. 6b. Not only the filtering effect improved but the overall regenerative gain has increased



**Noise and Stability in Modelocked Soliton Lasers, Figure 6**

Power gain of a regenerative laser amplifier as a function of detuning from the atomic line center. **a** Mirror reflectivities  $R_1 = R_2 = 30\%$ . Regenerative gain shown for 4 values of midband atomic gain. Broad “ATOMIC GAIN” curve corresponds to midband gain  $2\alpha_{m_0}p = 0.95$ . Internal round-trip loss coefficient  $2\alpha_{op} = 0.4$ . **b** Mirror reflectivities  $R_1 = R_2 = 60\%$ . Other parameters identical to **a**. (Adapted from [4])

**Noise and Stability in Modelocked Soliton Lasers, Table 1**

Parameters for regenerative laser amplifier with power gains plotted in Fig. 6. Additional assumptions,  $\lambda = 800$  nm, cavity length  $L = 1.5$  m, axial mode spacing  $\Delta\nu_{ax} = 100$  MHz

MIDBAND GAIN COEFF. $2\alpha_{m_0}p_m$	LOSS COEFF. $2\alpha_{op}$	MIDBAND PWR GAIN $\exp[2\alpha_{m_0}p_m]$	$R_{1,2}$	ROUND TRIP POWER GAIN $ \hat{g}_{rt} ^2$	REGEN. POWER GAIN	CAVITY LINE- WIDTH $\Delta\nu_{cav}$	FINESSE $\mathcal{F}$	Q
0	0.04	1	30%	0.087	0.964	45.09 MHz	2.413	$8.311 \times 10^6$
			60%	0.346	0.925	17.30 MHz	5.849	$2.167 \times 10^7$
0.6	0.04	1.822	30%	0.158	1.783	31.75 MHz	3.282	$1.180 \times 10^7$
			60%	0.630	4.983	7.375 MHz	13.58	$5.081 \times 10^7$
0.8	0.04	2.226	30%	0.192	2.274	27.84 MHz	3.707	$1.346 \times 10^7$
			60%	0.770	15.560	4.167 MHz	24.00	$8.992 \times 10^7$
0.95	0.04	2.586	30%	0.224	2.779	25.03 MHz	4.098	$1.498 \times 10^7$
			60%	0.894	85.54	1.776 MHz	56.27	$2.110 \times 10^8$

from a peak value of 2.8 to a peak of 86 (note the change in the vertical scale).

There are two conventional metrics for describing the relative linewidth of a resonator. In the case of a laser resonator where the wavelength is usually much smaller than the physical cavity dimensions, the *fineness*,  $\mathcal{F}$ , is the ratio of the axial mode spacing to the cavity linewidth

$$\mathcal{F} \equiv \frac{\Delta\omega_{ax}}{\Delta\omega_{cav}} = \frac{\Delta\nu_{ax}}{\Delta\nu_{cav}}. \quad (39)$$

This is the commonly used figure of merit for Fabry–Perot etalons [22]. The general concept of  $Q$  described earlier (35) can also be used to relate the cavity linewidth to the *center frequency* [4] according to

$$Q \equiv \frac{\omega_a}{\Delta\omega_{cav}} = \frac{\nu_a}{\Delta\nu_{cav}}. \quad (40)$$

From Table 1 we see that both the finesse and the  $Q$  improve by a factor of  $\approx 14$  when the mirror reflectivity is increased from 30% to 60% for a constant round-trip power gain of  $\exp[2\alpha_{m_0}p_m] = 2.586$ . This dramatic and nonlinear relationship between round-trip gain/loss and overall amplifier behavior hints at what happens as the round-trip gain approaches and just compensates for all of the round-trip losses. At that point, the regenerative gain goes to infinity, the linewidth goes to zero (in this model) and no light input is required. In fact, this is just the condition for the onset of laser oscillation. With no input applied to the regenerative amplifier, a single spontaneously emitted photon which happens to propagate exactly down the resonator axis is reflected (with probability equal to the mirror reflectivity) back to the gain medium, is amplified and returns again via the other mirror. This sets off the

avalanche of field buildup until the gain medium “saturates” its gain down to the point where the net round-trip gain exactly balances the round-trip loss. (Note: the saturation phenomenon occurs when the the optical field begins to deplete the population inversion faster than it can be maintained by the pumping rate). This is an important and fascinating component of lasers (indeed of all oscillator physics) but will be left out of the present discussion in the interest of brevity. The reader may consult any of the standard textbooks on laser physics or optical spectroscopy for a comprehensive treatment, e. g. [4,9,10].

### Laser Dynamics

Now that we have developed the basic concepts behind laser gain and cavity resonance, which together set the stage for optical amplification and finally oscillation, we must address the time dependence of the various coupled quantities. This is the domain of *laser dynamics* and in it we find answers to questions such as “how does a laser evolve from a cold cavity to oscillation upon being switched on?” and “what happens to an otherwise stable laser oscillator if it is subject to a sudden perturbation?” As pointed out in Sect. “Introduction”, these phenomena fall under the general heading of “large scale instabilities” and the reader can find more complete treatments in, e. g., [5,6,12,13,14,16,23]. For our purposes here, we will gather up the results of the analysis and apply it to a few of the myriad interesting (in)stability problems.

### Reduction of the Full Laser Equations

The major physical variables of a laser oscillator are the electric field  $E(t)$ , the atomic polarization (i. e. dipole moment per-unit-volume)  $P(t)$  and the energy level populations  $N_n(t)$ . Specifically, it is the *population difference*  $\Delta N = N_1(t) - N_2(t)$  that results in the gain or loss in the laser. The physics that couples the three variables is rather involved [4,5,6,12,16] and we will present only the main results here.

Upon solving Maxwell’s and Schrödinger’s equations for the motion of the electromagnetic fields and the atomic states of the atoms in the laser cavity, one obtains the following system of coupled nonlinear differential equations in “neoclassical” form [4],

$$\frac{d^2 E(t)}{dt^2} + \gamma_c \frac{dE(t)}{dt} + \omega_c^2 E(t) = -\frac{1}{\epsilon} \frac{d^2 P(t)}{dt^2} \quad (41)$$

$$\frac{d^2 P(t)}{dt^2} + \Delta\omega_a \frac{dP(t)}{dt} + \omega_a^2 P(t) = \kappa \Delta N(t) E(t) \quad (42)$$

$$\frac{d\Delta N(t)}{dt} + \frac{\Delta N(t) - \Delta N_0}{T_1} = -\frac{2^*}{\hbar\omega} E(t) \frac{dP(t)}{dt} \quad (43)$$

where  $\Delta N(t)$  is the atomic density difference averaged over the cavity volume  $V_c$

$$\Delta N(t) \equiv \frac{1}{V_c} \iiint \Delta N(\vec{r}, t) d\vec{r}, \quad (44)$$

$\Delta N_0$  is the equilibrium population density difference in the absence of driving fields,  $\gamma_c$  is the cavity field energy decay rate which takes into account losses as well as mirror output coupling, etc.,  $\omega_c$  is the cavity resonant frequency,  $T_1$  is the population-recovery or energy-relaxation time from level  $E_2 \rightarrow E_1$  (see p. 205 of [4] for an excellent discussion of this concept),  $2^*$  is the “bottlenecking factor” (that is, if the atoms get stuck emptying out of the lower energy level,  $2^* = 2$ , if not,  $2^* = 1$ , which constitutes a “good” laser) and, finally,

$$\kappa \equiv \frac{3^* \epsilon \lambda^3 \omega_a \gamma_{\text{rad}}}{4\pi^2} \eta \quad (45)$$

is the coupling constant linking the polarization to the product of the population difference and the electric field (i. e. the stimulated emission). Here  $\eta$  is a “filling factor” which expresses the fraction of the cavity mode volume filled by the active laser atoms [4] and the  $3^*$  describes the tensor coupling between the fields and the polarization ( $0 \leq 3^* \leq 3$ ).

This rather formidable-looking set of coupled Eqs. (41)–(43) can describe almost all dynamic laser phenomena but certain approximations can be made which allows us to simplify them without compromising rigor. The first approximation is based on the fact that both of the field amplitudes of interest will change slowly on a time scale compared to the optical carrier frequency. The field amplitudes in “phasor form” separate the slowly-varying complex envelope from the time-harmonic carrier,

$$E(t) = \frac{1}{2} [\tilde{E}(t)e^{i\omega t} + \tilde{E}^*(t)e^{-i\omega t}] \quad (46)$$

$$P(t) = \frac{1}{2} [\tilde{P}(t)e^{i\omega t} + \tilde{P}^*(t)e^{-i\omega t}]. \quad (47)$$

If we substitute these expressions into (41)–(43) and assume that the time variations of the phasor amplitudes along with the cavity decay rate  $\gamma_c$  and atomic linewidth  $\Delta\omega_a$  are all slow compared to the carrier frequency  $\omega$ , the atomic line center frequency  $\omega_a$  and the cavity resonance frequency  $\omega_c$ , the general Eqs. (41)–(43) become

$$\frac{d\tilde{E}(t)}{dt} + [\gamma_c/2 + i(\omega - \omega_c)] \tilde{E}(t) = -i\frac{\omega}{2\epsilon} \tilde{P}(t) \quad (48)$$



$$\frac{d\tilde{P}(t)}{dt} + [\Delta\omega_a/2 + i(\omega - \omega_a)] \tilde{P}(t) = -i\frac{\kappa}{2\omega} \Delta N(t) \tilde{E}(t) \quad (49)$$

$$\begin{aligned} \frac{d\Delta N(t)}{dt} + \frac{\Delta N(t) - \Delta N_0}{T_1} \\ = i\frac{2^*}{4\hbar} [\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)] . \end{aligned} \quad (50)$$

The steps leading to the above equations are commonly known as the *slowly-varying envelope approximation* (SVEA) since we have replaced equations describing the dynamics of instantaneous quantities with equations describing the evolution of the much more slowly-varying envelope functions.

### Coupled Cavity-Atom Rate Equations

At this point, we could integrate the coupled Eqs. (48)–(50), but there are several other simplifications we can take that help lend considerable physical insight into laser dynamics and yet preserve the essential physics.

(Note: An early description of the dynamics of the photon and population densities in a laser was given by Dunsmuir [24]. The description is in terms of rate equations which can be derived from the full set of coupled equations above and understood from elementary energy-balance considerations.)

**Electric Field in Terms of Photons** The first step is to write the electric field in terms of quanta of electromagnetic energy, or photons. The number of photons in the cavity is given by the total electromagnetic energy divided by  $\hbar\omega$ , the energy per photon, or

$$n(t) = \frac{\epsilon V_c}{2\hbar\omega} |\tilde{E}(t)|^2 = \frac{\epsilon V_c}{2\hbar\omega} \tilde{E}(t)\tilde{E}^*(t) . \quad (51)$$

Differentiating this and substituting from (48) for  $d\tilde{E}(t)/dt$  and  $d\tilde{E}^*(t)/dt$  results in

$$\frac{dn(t)}{dt} + \gamma_c n(t) = i\frac{V_c}{4\hbar} [\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)] . \quad (52)$$

**Population Density to Total Population** In order to make the cavity-averaged population density equation have a similar driving term as the photon equation, we multiply the distributed density by the cavity mode volume to get the total population difference;  $\Delta N(t) = V_c \Delta N(t)$ . Applying this to (50) we have

$$\begin{aligned} \frac{d\Delta N(t)}{dt} + \frac{\Delta N(t) - \Delta N_0}{T_1} \\ = -i\frac{2^* V_c}{4\hbar} [\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)] . \end{aligned} \quad (53)$$

**Adiabatic Elimination of the Polarization Field** In many cases of interest, the dephasing of the polarization field  $P(t)$  happens on a time scale that is much faster than the atomic energy decay time  $\gamma^{-1}$  and, therefore, the polarization field's “phase memory” is continuously interrupted making it possible for the envelope to follow fluctuations in the driving source term  $\Delta N(t)\tilde{E}(t)$  in (49). This allows us to establish a linear relationship between the phasor amplitudes of the electric field and the polarization with essentially no time delay. We consider the polarization to adiabatically follow the electric field and thus can solve the SVEA equation (49) in the steady state ( $d\tilde{P}(t)/dt = 0$ ) to obtain

$$\tilde{P}(t) \approx \frac{-i\kappa}{\omega\Delta\omega_a} \frac{1}{1 + 2i(\omega - \omega_a)/\Delta\omega_a} \Delta N(t)\tilde{E}(t) . \quad (54)$$

This is frequently called the “linear susceptibility” or “rate equation” approximation. We can now use this result on the right-hand sides of (52) and (53),

$$\begin{aligned} i\frac{V_c}{4\hbar} [\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)] \\ = -\frac{\kappa}{2\hbar\omega\Delta\omega_a} \times \frac{\Delta N(t)}{1 + [2(\omega - \omega_a)/\Delta\omega_a]^2} |\tilde{E}(t)|^2 . \end{aligned} \quad (55)$$

Substituting in for the coupling constant  $\kappa$  (45) and the electric field in terms of the photon number (52), we can write

$$\begin{aligned} i\frac{V_c}{4\hbar} [\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)] = \\ -\frac{3^* \omega_a \gamma_{\text{rad}} \eta \lambda^3}{4\pi^2 \Delta\omega_a V_c} \times \frac{1}{1 + [2(\omega - \omega_a)/\Delta\omega_a]^2} \Delta N(t)n(t) \end{aligned} \quad (56)$$

$$= -K\Delta N(t)n(t) \quad (57)$$

where we have defined a new “rate-equation coupling constant”  $K$ . Substituting (57) into (52) and (53), we have

$$\frac{dn(t)}{dt} + \gamma_c n(t) = -K\Delta N(t)n(t) \quad (58)$$

and

$$\frac{d\Delta N(t)}{dt} + \frac{\Delta N(t) - \Delta N_0}{T_1} = -2^* K\Delta N(t)n(t) . \quad (59)$$

These equations describe the balance between population levels and photon number within the cavity for a two-level system. In order to predict dynamic laser phenomena, especially large transients, we need to add a couple of

more terms. First of all, most lasers rely on a four-level energy scheme (see Fig. 3). The lowest level,  $E_0$  acts as a reservoir of atoms while the top level,  $E_3$  is a momentary stop on the way to the metastable level  $E_2$ . In a typical 4-level “good” laser, pumping from the ground state  $E_0$  to level  $E_3$  quickly relaxes into level  $E_2$  which is long-lived and allows for population inversion with respect to level  $E_1$ . Level  $E_1$  will also empty quickly into the ground-state level and be ready for pumping back up into  $E_2$  via  $E_3$ . With this in mind, it is a good approximation to set the level populations  $N_3 = N_1 = 0$ .

When a laser is actually oscillating, there must be a population inversion so that  $\Delta N(t) = N_1(t) - N_2(t) < 0$ . We can recast our equations in terms of this inversion by simply defining

$$N(t) \equiv N_2(t) - N_1(t) = -\Delta N(t). \quad (60)$$

Next, to account for the effect of pumping atoms from the lower laser level to the upper level to create the inversion, we will also add a term  $R_p(t)$  to (59). Also, for lasers operating in the visible and near-infrared portion of the spectrum the Boltzmann distribution governing the population levels in thermal equilibrium (i. e. with no electromagnetic signals driving the transitions) gives

$$\frac{N_2}{N_1} = \exp\left(-\frac{\hbar\omega_a}{kT}\right) \approx 0 \quad (61)$$

and thus  $\Delta N_0 = N_{10} - N_{20} \approx N_{10}$ . In a “good” four-level laser, there will be very fast relaxation out of the lower laser level always and therefore, to a good approximation  $N_{10} = 0$  and so  $\Delta N_0 = 0$ . Finally, since the lower level is assumed to empty very fast, there is no bottlenecking and  $2^* = 1$ . We can now rewrite (58) and (59) in simplified form,

$$\frac{dn(t)}{dt} = KN(t)[n(t) + n_{sp}] - \gamma_c n(t) \quad (62)$$

$$\frac{dN(t)}{dt} = R_p(t) - \gamma_2 N(t) - KN(t)n(t). \quad (63)$$

Notice that we have included one more term,  $n_{sp}$ , to account for photons spontaneously emitted into the desired cavity mode. This term is essential for getting the laser “started”, but once started, the total number of photons in the cavity mode,  $n(t)$ , far exceeds  $n_{sp}$ , as we shall see.

The coupled rate equations (62), (63) form an intuitive and accurate platform for understanding the dynamics between the pump source, the population inversion and the photons in a laser in the approximation that these processes occur on a time scale that is long compared to a cycle of the optical field. This is a reasonable assumption in almost all cases of interest.

A physical picture that is described by (62), (63) can be constructed by realizing that these equations form an energy storage and dissipation system. Energy is pumped into the atoms via  $R_p$  and released into the electromagnetic field by stimulated emission ( $KNn$ ). As the field builds up in the resonator, more energy is pulled out of the atoms and released through the output coupling mirror ( $-\gamma_c n$ ). Soon the atoms are somewhat depleted ( $-KNn$ ) and must “recharge” from the pump source  $R_p$ . The energy stored in the fields is reduced while the pump source increases the energy stored in the atoms. This “dance” between the energy stored in the atoms and the energy stored in the field is at the heart of the elementary “relaxation oscillations” and is also the first important dynamic laser phenomenon. Unfortunately, since the coupled rate equations (62), (63) are nonlinear, they cannot be solved in closed form. So, to get a more quantitative feeling for how the relevant quantities behave, let us first look at a couple of limiting cases.

**Steady-State Response** For a constant pumping rate  $R_p(t) = R_{p_0}$ , the laser will achieve a constant or steady-state operating point such that

$$\frac{dn(t)}{dt} = \frac{dN(t)}{dt} = 0. \quad (64)$$

Applying this to (62), (63) results in the following pair of equations

$$0 = KNn - \gamma_c n \quad (65)$$

$$0 = R_{p_0} - KNn - \gamma_2 N. \quad (66)$$

Notice that we have removed the time dependence and also the spontaneous emission term. Its role only becomes significant during the startup of dynamic processes.

There are two solutions to the system above. We see by inspection of (65) that either  $n = 0$  or  $N = \gamma_c/K$ . Let’s investigate both cases.

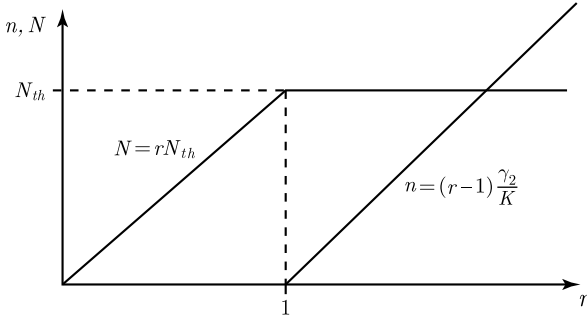
**Case 1:  $n = 0$**  Applying this condition to Eq. (66) yields

$$N = R_{p_0}/\gamma_2. \quad (67)$$

The population inversion varies directly with the pumping rate. Since there are no photons generated, we call this the “below threshold” state.

**Case 2:  $N = \gamma_c/K$**  Again we apply this condition to Eq. (66) and find

$$n = \frac{R_{p_0}}{\gamma_c} - \frac{\gamma_2}{K} = \left(\frac{R_{p_0}K}{\gamma_2\gamma_c} - 1\right) \frac{\gamma_2}{K}. \quad (68)$$



**Noise and Stability in Modelocked Soliton Lasers, Figure 7**

Photon number  $n$  and population inversion  $N$  as a function of the normalized pumping rate  $r = R_{P_0}/\gamma_2 N_{th}$ . At the first threshold,  $r = 1$ , ( $R_{P_0} = \gamma_2 N_{th}$ ), the population inversion is clamped at  $N = N_{th}$  and the photon number starts to rise linearly

We have deliberately written the result in this form to make a point. In *Case 2*, the population inversion is constant and is said to be “clamped” at the threshold level for lasing. We will call this the “threshold inversion density”

$$N(t) = N_{th} = \frac{\gamma_c}{K}. \quad (69)$$

The pumping rate which just achieves threshold inversion, and for which the photon number  $n$  begins to rise from zero, will be called the “first threshold” pumping rate;  $R_{P_{01}} = \gamma_2 N_{th}$ . Above this point, the photon density is linearly proportional to the pumping rate

$$n = n_{ss} = \left( \frac{R_{P_0}}{\gamma_2 N_{th}} - 1 \right) \frac{\gamma_2}{K} = (r - 1) \frac{\gamma_2}{K}. \quad (70)$$

This is the steady-state photon density and we have defined a “normalized pumping rate”,

$$r \equiv R_{P_0}/\gamma_2 N_{th}, \quad (71)$$

valid above the first threshold. Inserting typical numbers for a Ti:sapphire laser;  $\gamma_c = 1.280 \times 10^7 \text{ s}^{-1}$ ,  $\gamma_2 = 3.125 \times 10^5 \text{ s}^{-1}$  and  $K = 6.860 \times 10^{-17} \text{ m}^3 \text{ s}^{-1}$ , we find

$$N_{th} = 1.866 \times 10^{23} \text{ m}^{-3}, \quad n_{ss} = 4.555 \times 10^{21} \text{ m}^{-3} \quad (72)$$

when pumped at twice threshold ( $r = 2$ ).

**Step Response with no Spontaneous Emission** It is interesting to study the response of the system assuming no spontaneous emission into the cavity mode ( $n_{sp} = 0$ ). In this case, there are no photons to seed the stimulated emission process and therefore  $n(t) = 0$  for all time. The population inversion, however, is immediately driven by a step

function pumping rate  $R_P(t) = R_{P_0} H(t)$  where  $H(t)$  is the Heaviside step function, defined according to

$$H(t) = \begin{cases} 0, & t \leq 0 \\ 1, & t > 0. \end{cases} \quad (73)$$

With no photons to build up and stimulate downward transitions, Eq. (63) becomes

$$\frac{dN_o(t)}{dt} = R_P(t) - \gamma_2 N_o(t) \quad (74)$$

with solution

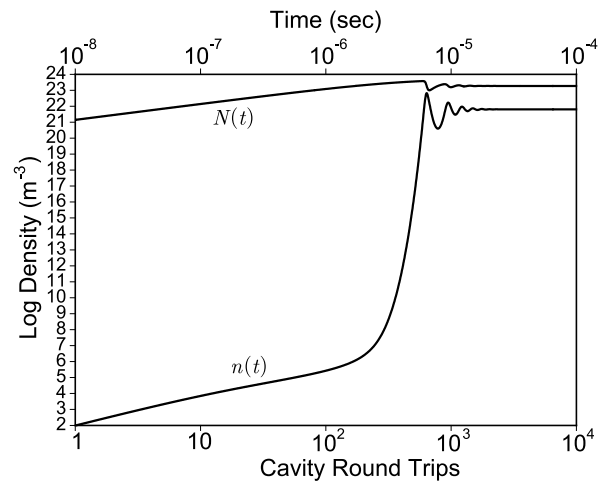
$$N_o(t) = \frac{R_{P_0}}{\gamma_2} [1 - e^{-\gamma_2 t}], \quad t > 0. \quad (75)$$

This tells us that the population inversion density rockets upward with a time constant  $\gamma_2$ . In the asymptotic limit, we find  $N_o(t \rightarrow \infty) = R_{P_0}/\gamma_2$ . Substituting from (70) we find that (75) can be written in the especially simple form

$$N_o(t) = r N_{th} [1 - e^{-\gamma_2 t}], \quad t > 0. \quad (76)$$

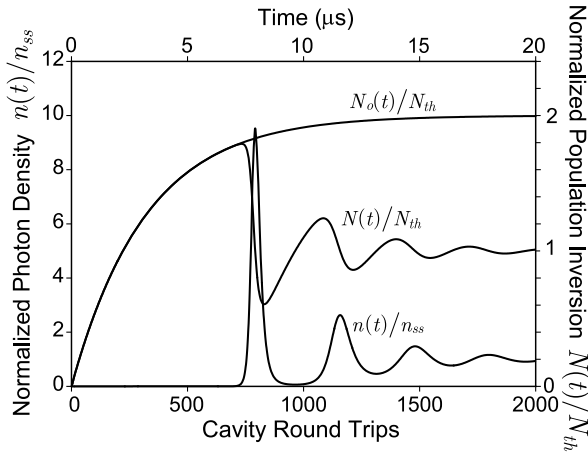
This says that in the absence of spontaneous emission photons to seed the stimulated growth process, the population inversion will grow smoothly to  $r$  times the threshold inversion obtained under actual lasing conditions.

**Full Step Response with Stimulated Emission** Consider now the solutions to the full coupled nonlinear rate equations (62), (63). These have been computed numerically for a typical 1.5 meter-long Ti:sapphire laser cavity



**Noise and Stability in Modelocked Soliton Lasers, Figure 8**

Laser photon density  $n(t)$  and population inversion  $N(t)$  for a step function pump  $R_P(t)$ . The normalized pumping rate  $r = 2.0$

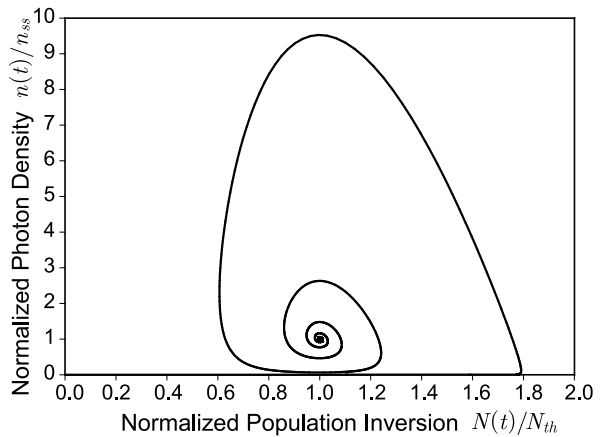


**Noise and Stability in Modelocked Soliton Lasers, Figure 9**  
Laser photon density  $n(t)$  and population inversion  $N(t)$  normalized to steady-state values for a step function pump  $R_p(t)$ . The normalized pumping rate  $r = 2.0$

and appear in Fig. 8 on a log-log scale. The spontaneous emission term,  $n_{sp}$ , of one photon-per-cavity-mode volume, stimulates growth in the cavity photon density,  $n(t)$  which climbs at a rate of about an order of magnitude per decade of round trips until  $n \approx n_{sp}$  and the photon density rapidly climbs by 15 orders of magnitude until the inversion density starts to get “eaten up” by the large number of photons. This can be seen more clearly if we look at a plot of the dynamics on a linear scale as in Fig. 9. At about 700 cavity round trips, the photon density starts to eat up the inversion and thus the gain. The inversion rapidly falls below threshold,  $N_{th}$  so that there is more net loss than gain in the system. This causes the photon number to fall rapidly below the steady-state value  $n_{ss}$  which in turn allows the inversion to climb back up. When the inversion rises above the threshold value, net gain is restored to the system and the photon number starts to climb again. This behavior, including the roughly  $90^\circ$  phase lag between photon density and population inversion, is characteristic of the standard “predator-prey” relations of the Lotka–Volterra model [25]. The difference here is that the oscillations of the dependent variables die away with a time constant  $\tau \approx \gamma_2^{-1}$ .

The top curve in Fig. 9 shows the response of the population inversion density,  $N_o(t)$ , in the absence of the spontaneous emission factor  $n_{sp}$ . This is the solution presented in (76). There is no accompanying photon growth to eat away at the population inversion and therefore there are no oscillations and the population grows asymptotically to  $rN_{th}$ . In the case depicted here,  $r = 2.0$ .

A common format for viewing the dynamic behavior of two variables coupled by nonlinear differential equa-



**Noise and Stability in Modelocked Soliton Lasers, Figure 10**  
Laser photon density  $n(t)$  and population inversion  $N(t)$  for a step function pump  $R_p(t)$  displayed in phase space

tions with an independent variable (or parameter) such as time, is the phase plane [26]. In this format, one variable is plotted against the other with time as a parameter. Thus curves or trajectories in the phase plane represent the evolution of the system. The solution to the laser startup problem shown in Fig. 9 is plotted in the phase plane portrait of Fig. 10. Both the photon density and the population eventually settle into their steady-state values which appears as a trajectory circling around and finally settling into a “focal point”. The phase plane is a highly useful device for exhibiting a wide range of nonlinear behavior and can often reveal subtle features of a nonlinear system not easily observed when plotting either of the dynamical variables alone as a function of time.

### Laser Instability, Chaos and the Lorenz Equations

The general linear trend of output power versus pumping rate shown in Fig. 7 and even the transient responses depicted in Figs. 8–10 represent solutions to well-behaved “stable” equations. Gentle changes in any of the system parameters will result in gentle changes in output power once the system’s transient response has died out. In the laboratory, however, wild and chaotic behavior occasionally occurs in a laser system and although rather undesirable, analysis to try and explain those aberrant phenomena was slow in developing. One of the early papers on the subject was by Kaplan and Zier [27] who derived a set of equations for the steady-state behavior of a 3-level “optical maser” and compared linearized solutions against data from Bostick and O’Connor [28]. Tang questioned the validity of the rate equation approach that was commonly used at the time for optical masers and discovered that paying close

attention to the relevant time constants (for energy decay, polarization dephasing and cavity lifetime), could admit terms into more complete equations that would give rise to undamped transient solutions [29]. In general, these treatments only revealed broad trends of instability without producing a very coherent picture of the whole landscape of the phenomena. It would not be until years after the birth of the laser that results from a completely unrelated field were recognized as being important for laser physics in the context of shedding new and more comprehensive light on the subject.

In 1963, Edward N. Lorenz published a seminal paper on the topic of hydrodynamic flow in which coupled nonlinear differential equations were developed to predict the flow in a forced dissipative system [30]. He discovered that the system would admit nonperiodic solutions which had very sensitive dependence on their initial conditions and often would yield highly irregular results. This work set the seeds for the explosive growth of the field of chaos science years later [31,32].

The equations formulated by Lorenz describe the flow of fluid in a Bénard cell which is a rectangular tank heated on the bottom [32]. At low temperature differences between the top and bottom, a steady-state gradient is produced since the liquid viscosity is high enough to prevent convection. Above a critical temperature, however, the fluid starts to move in counter-rotating rolls (convection) and at still higher temperatures turbulence and chaos sets in. Lorenz's equations, in normalized form, are

$$\frac{dx}{dt} = \sigma(y-x), \quad \frac{dy}{dt} = x(r-z)-y, \quad \frac{dz}{dt} = xy-bz. \quad (77)$$

For the fluid problem, in these equations  $x$  is proportional to the intensity of the convective motion,  $y$  is proportional to the temperature difference between the ascending and descending currents and  $z$  is proportional to the degree to which the temperature profile deviates from linearity in the vertical direction. The other terms are de-

fined in Table 2. The Prandtl number is the ratio of viscous to thermal diffusivity.  $R$  is the Reynolds number and  $R_c$  is the Reynolds number for which the flow changes from conduction to convection.  $k_1$  is a dimensionless wavenumber.

In 1975, Haken [33] pointed out an analogy between the the equations Lorenz had developed and those governing the evolution of the laser electric field, polarization and population inversion in the slowly-varying envelope approximation. He suggested that, based on the close analogy to the Lorenz equations, new instabilities in lasers should be found when operated in special regions of parameter space. This gave birth to new analytical and theoretical studies of laser instabilities, some of which which were previously known, but not well understood. Caspersen [34] was also intrigued by the problem of spontaneous pulsations and instabilities in lasers and independently developed a quantum mechanically-based set of equations which predicted this type of behavior. Shortly thereafter, many papers began to appear in the literature based on various version of the coupled equations (48)–(50) and the field of chaotic laser behavior was off and running [35,36,37,38,39,40,41,42,43,44,45,46,47,48].

The analogy between the coupled field-atom SVEA equations (48)–(50) and the Lorenz equations (77) is intuitively clear upon visual comparison. However, in order to make the Lorenz equations analytically useful, one must specify the exact transformations between the envelope and population variables and the variables of the hydrodynamic system. This will depend, of course, on the exact form of the laser equations and these will vary depending on the laser configuration and the author's approach to the analysis. Thus, when studying the literature, one may come across a myriad of differing definitions for the Lorenz variables in terms of the laser variables. For our problem in the present context, we make the following associations between the dependent variables in the two systems;

$$x = \sigma \tau_c \sqrt{\frac{2^* \kappa}{\hbar \omega}} \tilde{E}, \quad y = \sigma \tau_c \sqrt{\frac{2^* \kappa}{\hbar \omega}} \left( -i \frac{\omega \tau_c}{\epsilon} \right) \tilde{P}, \\ z = \frac{\kappa \sigma \tau_c^2}{\epsilon} [\Delta \mathcal{N} - \Delta \mathcal{N}_0] \quad (78)$$

along with rescaling the time coordinate  $t' \rightarrow t/T_2$ . When these expressions, and those for the coefficients  $\alpha$ ,  $r$  and  $\beta$  (Table 2) are substituted into the Lorenz equations (77), and the derivatives are changed accordingly ( $d/dt \rightarrow (1/T_2) d/dt'$ ), the full set of laser equations (48)–(50) are recovered for the “on resonance” condition,  $\omega = \omega_a = \omega_c$ . Within this context, we may make the following asso-

**Noise and Stability in Modelocked Soliton Lasers, Table 2**  
Relationships between coefficients of Lorenz equations for hydrodynamic flow and the laser equations in the slowly-varying envelope approximation

Coefficient	Bénard convection	Laser
$\sigma$	Prandtl number	$\gamma_c T_2/2 = T_2/2\tau_c = \gamma_c/2\gamma_\perp$
$r$	$R/R_c$	$-\frac{\kappa T_2 \tau_c}{2\epsilon} \Delta \mathcal{N}_0$
$b$	$\frac{4\pi^2}{\pi^2 + k_1^2}$	$\gamma T_2 = T_2/T_1 = \gamma_\parallel/\gamma_\perp$



ciations;

$$x = \text{FIELD AMPLITUDE} \quad (79)$$

$$y = \text{POLARIZATION} \quad (80)$$

$$r - z = \text{POPULATION INVERSION} . \quad (81)$$

We may also associate with  $r$  an effective pumping rate since it sets the population levels in the absence of driving fields. The laser equations in the SVEA written in the form of the Lorenz equations are sometimes called the “Lorenz–Haken” equations or “semi-classical Maxwell–Bloch” equations.

What can we learn about laser dynamics from the Lorenz equations? A good deal, as it turns out. Theoretical predictions and experimental verifications in a large number of laser systems have been made over the years and several excellent books covering many aspects of general laser dynamics, including this analysis, have appeared [5,6,12,16,49]. However, in the limited scope of this chapter, we will only have space to point out the highlights of the varied phenomena described by (77).

First off, we remark that within the simple set of Eqs. (77), there are only three dependent variables, one independent variable and three parameters. These means that the time evolution of  $x(t)$ ,  $y(t)$  and  $z(t)$  depends critically on the values of the three parameters  $r$ ,  $\sigma$  and  $b$ . Let us now undertake a brief study of the general behavior of the Lorenz equations for some specific values of the parameters. We begin by examining the stationary or steady-state

solutions of the Lorenz equations, whence we obtain

$$\frac{dx}{dt} = 0 = -\sigma(x - y) \quad (82)$$

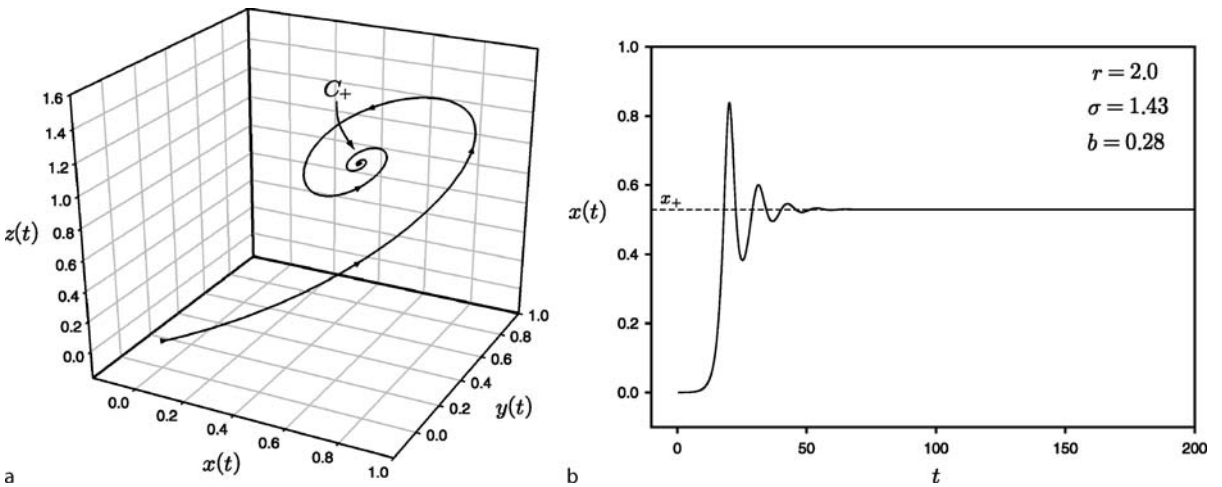
$$\frac{dy}{dt} = 0 = (r - z)x - y \quad (83)$$

$$\frac{dz}{dt} = 0 = xy - bz \quad (84)$$

where, for simplicity, we have dropped the prime notation on the time variable ( $t' \rightarrow t$ ). By inspection, Eq. (82) implies  $x = y$  and when this is substituted into (83) we have  $z = r - 1$ . Finally, substituting this result into (84), we find that both  $x$  and  $y$  become  $(x, y) = \pm\sqrt{bz} = \pm\sqrt{b(r-1)}$ . Summarizing, our steady-state solutions can be written

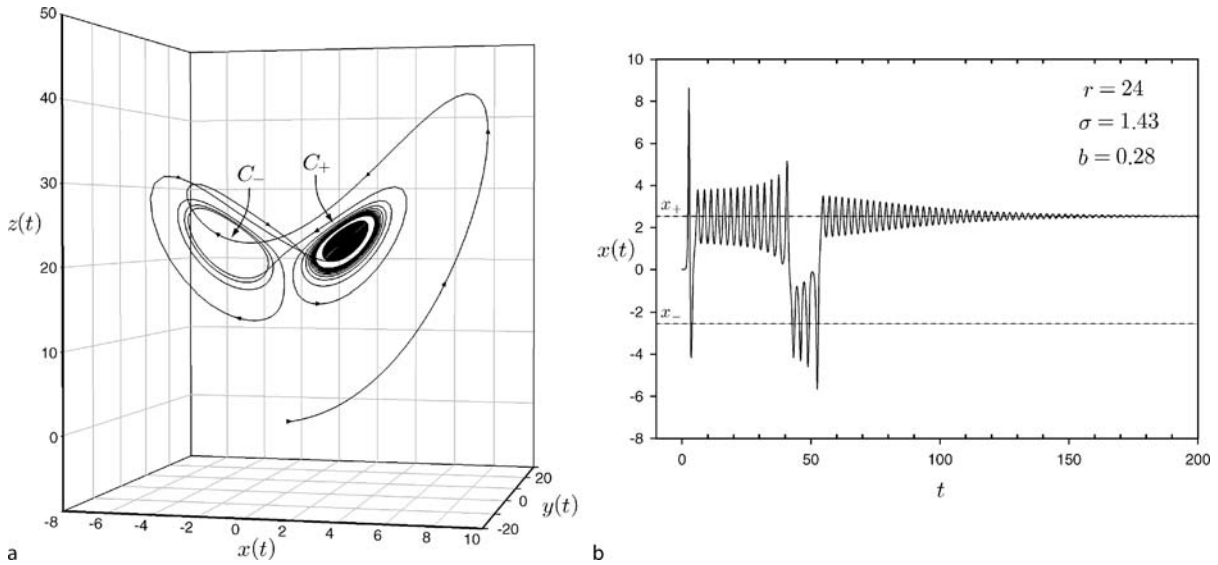
$$C_{\pm} = \begin{bmatrix} x_{\pm} \\ y_{\pm} \\ z \end{bmatrix} = \begin{bmatrix} \pm\sqrt{b(r-1)} \\ \pm\sqrt{b(r-1)} \\ r-1 \end{bmatrix} . \quad (85)$$

Let us examine the behavior of solutions as a function of the parameter  $r$  [16]. For  $0 \leq r \leq 1$ , the only physically meaningful solutions require that  $x = y = z = 0$ . This corresponds to the laser being pumped below threshold. There is not enough gain in the system to overcome the losses and consequently there is no buildup of the cavity fields or polarization. However, as we saw in the simplified model in Subsect. “Coupled Cavity-Atom Rate Equations”, the below-threshold operation results in a linear increase in the population inversion (see Fig. 7).



**Noise and Stability in Modelocked Soliton Lasers, Figure 11**

**a** Parametric plot of the solutions  $x(t)$ ,  $y(t)$ ,  $z(t)$  of the Lorenz equations for  $r = 2.0$ ,  $\sigma = 1.43$  and  $b = 0.28$ . Coordinate of focus  $C_+$  given by (85). **b**  $x(t)$  alone (electric field) showing damped response relaxing to the steady state solution  $x_+ = +0.53$



**Noise and Stability in Modelocked Soliton Lasers, Figure 12**

**a** Parametric plot of the solutions  $x(t)$ ,  $y(t)$ ,  $z(t)$  of the Lorenz equations for  $r = 24.0$ ,  $\sigma = 1.43$  and  $b = 0.28$ . Coordinates of foci  $C_{\pm}$  given by (85). **b**  $x(t)$  alone (electric field) showing initially erratic response finally relaxing to damped steady state solution  $x_+ = +2.54$

For  $1 \leq r \leq 35.850$  the fields and population inversion rise to one of the two stable solutions in a uniform manner as shown in Fig. 11. A visually interesting way of plotting the results is shown in Fig. 11a where the three dependent variables,  $x(t)$ ,  $y(t)$  and  $z(t)$  are plotted parametrically as a function of time. For this figure and the following, we have solved the Lorenz equations using  $\sigma = 1.43$  and  $b = 0.28$ . We see that the solutions asymptotically spiral into the stable point  $C_+$ . Figure 11b shows just the function  $x(t)$  (electric field) which has the traditional damped relaxation oscillation behavior as it approaches the stable value  $x_+ = +0.53$ .

As we turn up the value of  $r$ , we no longer find that the solutions uniformly converge onto one of the stable points,  $C_{\pm}$ . Rather, they tend to circle around the stable points in phase space and eventually land on one of the two. An example of this phenomenon is shown in Fig. 12 which was computed for  $r = 24$ . Notice that the solution trajectory (Fig. 12a) initially makes one quick pass around  $C_+$  followed by a quick loop around  $C_-$  and then back to  $C_+$  where it circles another 13 times, loops back to  $C_-$  and then finally goes into a long decaying orbit around  $C_+$ . Figure 12b again shows just the  $x(t)$  solution and we can easily see how the electric field bounces back and forth between the two stable values  $x_{\pm} = \pm 2.54$  until it settles down to  $x_+$ . This seemingly bizarre behavior is characteristic of the Lorenz equations and depends critically on the value of  $r$ . At special critical values (e.g.  $r = 1, 4.9, 8$ ,

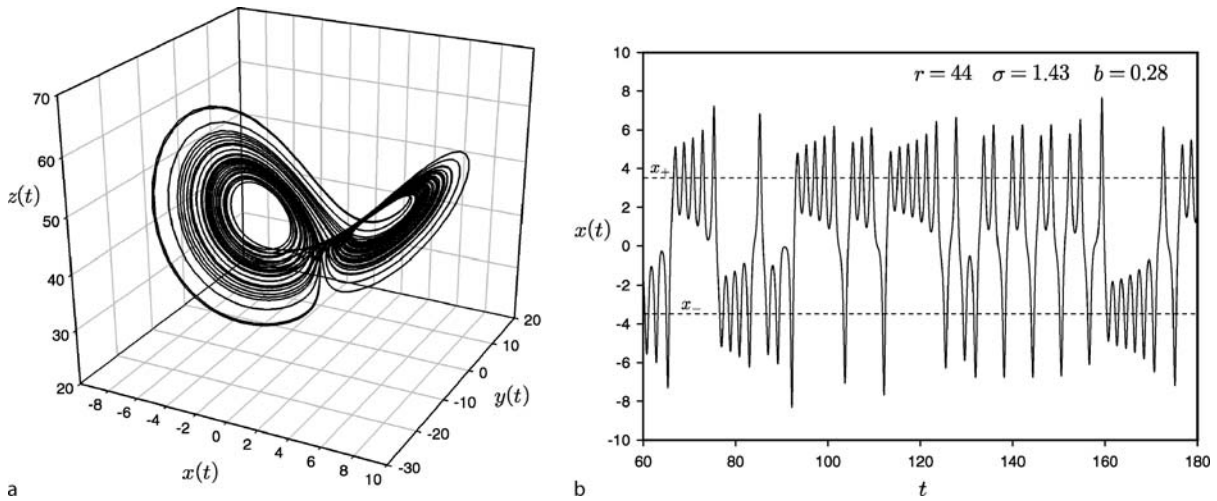
etc.), there are pronounced changes in the behavior of the orbits in phase space and these points are referred to as *bifurcations*.

When  $r$  is further increased beyond the value  $r = 35.83$ , a curious thing happens. The trajectory in phase space no longer settles into one of the two stable solutions,  $C_{\pm}$ , but instead circles about them indefinitely, never repeating any orbit and thus never becoming a “stable” solution. Figure 13 shows an example of this behavior for  $r = 44$ . In the parametric plot of Fig. 13a the initial rise from the origin has been excluded. In Fig. 13b we see the aperiodic, erratic behavior which constitutes “chaos”. The region of phase space surrounding the so-called “steady-state” solutions,  $C_{\pm}$ , about which the trajectory propagates is called an *attractor*, presumably in the sense that a gravitational potential holds an orbiting body within its realm. When the dimensionality of the attractor is non-integer (i.e. fractal), then it is called a *strange attractor* [6,16].

If we continue turning up  $r$ , we find that the laser output becomes unstable at the value [16]

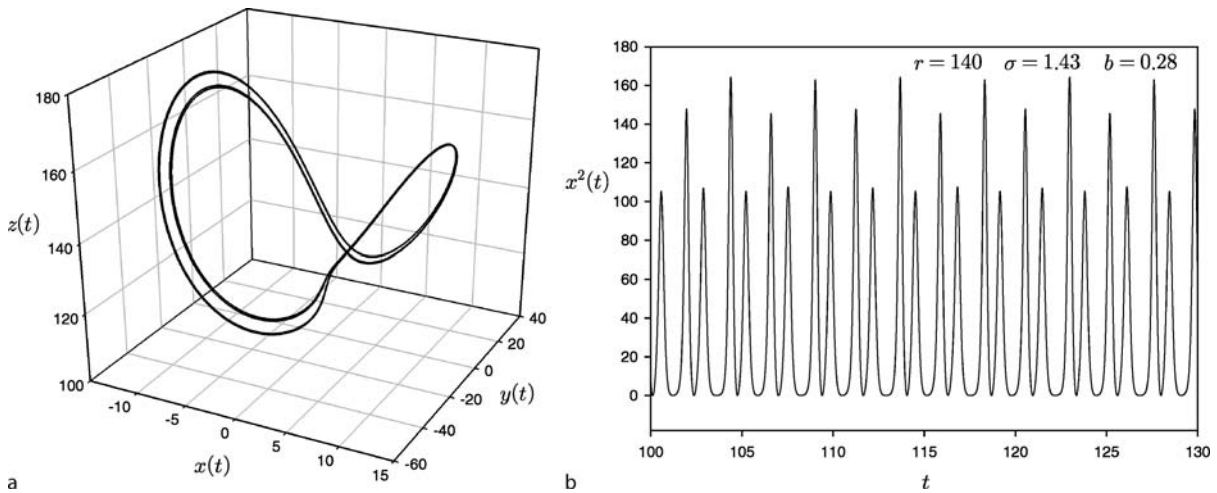
$$r = r_H = \sigma(\sigma + b + 3)/(\sigma - b - 1) \quad (86)$$

which for the examples shown above occurs at  $r = 44.90$ . Physically, the potential well at the foci of the strange attractors has gone to zero depth. This point is referred to as the “second laser threshold”. A necessary condition for



**Noise and Stability in Modelocked Soliton Lasers, Figure 13**

**a** Parametric plot of the solutions  $x(t), y(t), z(t)$  of the Lorenz equations for  $r = 44.0, \sigma = 1.43$  and  $b = 0.28$ . Solution propagates around foci ("strange attractors"). Coordinates of foci  $C_{\pm}$  given by (85). **b**  $x(t)$  alone (electric field) showing chaotic behavior with no repeating patterns



**Noise and Stability in Modelocked Soliton Lasers, Figure 14**

**a** Parametric plot of the solutions  $x(t), y(t), z(t)$  of the Lorenz equations for  $r = 140.0, \sigma = 1.43$  and  $b = 0.28$ . Notice that the orbit only close upon itself after 4 traverses around the original stable points  $C_{\pm}$ , thus demonstrating a four-fold period multiplication. **b**  $x^2(t)$  alone (proportional to power in electric field)

this to occur can be seen from (86),

$$\sigma > b + 1, \quad \text{or} \quad \rightarrow \quad \tau_c < \frac{1}{T_1^{-1} + T_2^{-1}}. \quad (87)$$

This means that the cavity decay rate must exceed the sum of the energy decay plus the polarization decay rates. This is known as the *bad cavity condition* and in general, it is hard to achieve, especially in the visible portion of the spectrum. However, the spontaneous emission rate (which is, effectively,  $T_1^{-1}$ ) is proportional to

$\omega_a^2$  (6) which favors long wavelength lasers. Also, for gas lasers the polarization decay rate is dominated by collisions which is proportional to pressure. These two facts favor long-wavelength gas lasers as a proving ground for investigating many of the predictions of the Lorenz equations and, indeed, most of the early work in the field was conducted with, e. g., laser-pumped ammonia ( $\text{NH}_3$ ) ring lasers ( $\lambda = 81 \mu\text{m}$ , homogeneously broadened) [16] and discharge-pumped He-Xe lasers ( $\lambda = 3.5 \mu\text{m}$ , inhomogeneously broadened) [34].

**Noise and Stability in Modelocked Soliton Lasers, Table 3**  
Range of values of pump parameter  $r$  and related laser behavior

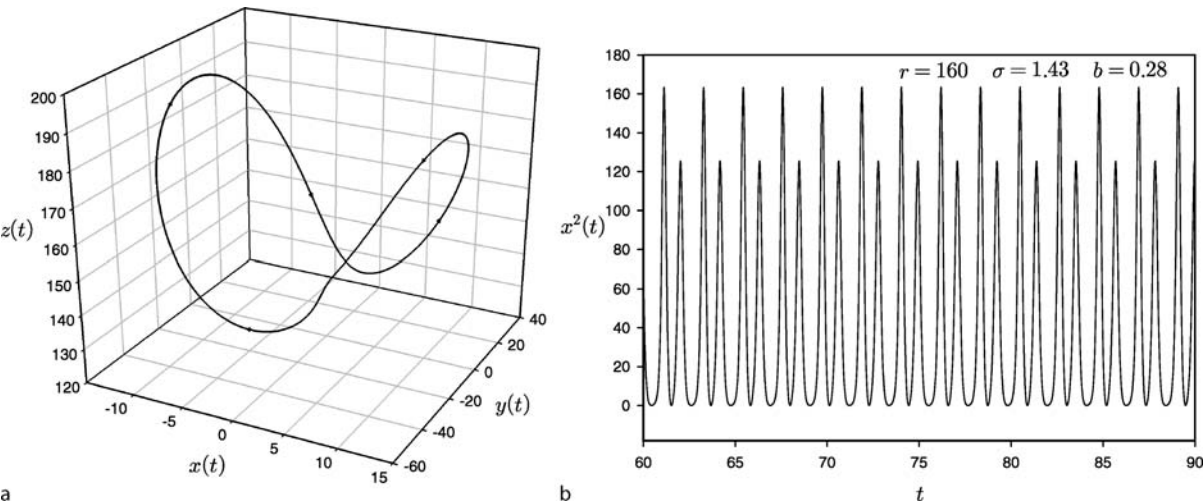
$r$	CHARACTERISTIC FEATURES
$0 \leq r \leq 1$	Below threshold. Spontaneous emission.
$r = 1$	First laser threshold.
$1 \leq r \leq r_A$	Asymptotically approaches stable solutions $C_{\pm}$ .
$r = r_A$	$r_A = 35.83$ . Upper limit of asymptotic relaxation to $C_{\pm}$ .
$r > r_A$	No periodicity. Irregular and chaotic. Strange attractors.
$r_H$	Second laser threshold. $r_H = \sigma(\sigma + b + 3)/(\sigma - b - 1) = 44.9$
$r > r_H$	Unstable.
$r = r_B$	$r_B \approx 138$ . Boundary for strange attractor. Orbits start to close.
$r > r_C$	$r_C \approx 140$ . Closed orbits. Period multiplication.

Continuing to increase  $r$  demonstrates another interesting phenomenon, that of period multiplication. In Fig. 14, we have plotted the solutions for  $r = 140$  and see that the trajectory in phase space is almost a closed curve but, in fact, cycles 4 times before repeating. This is a “period-quadrupling” ( $T/4$ ) bifurcation and occurs throughout the region for various values of  $r \gtrsim 140$ . Finally, as we reach far beyond  $r = 150$ , we start to see solid closed orbits in phase space (Fig. 15). A short summary of the important ranges of the pumping parameter  $r$  and the related laser characteristics is given in Table 3 where we have borrowed some nomenclature from the excellent text by Weiss and Vilaseca [16].

Now, we have only given a rather brief account of these interesting phenomena within the context of one specific set of laser parameters. There are many other laser systems that exhibit unusual chaotic behavior, such as semiconductor diode lasers. The reader interested in a broader perspective is encouraged to see, for example, [23,50].

### Modelocking

All of the principles discussed so far regarding lasers have related to continuous-wave (CW) operation. Lasers can also be configured to send out pulses with a tremendous amount of energy and short duration. The future (and potential) of inertial confinement fusion rests on this very fact. There are predominantly two different techniques for producing short pulses in lasers. In the first, called Q-switching, the atoms in the laser gain medium are pumped up to an inverted condition far beyond that normally required to achieve CW oscillation. This can happen if an intracavity loss mechanism is introduced to prevent the feedback which initiates and sustains lasing. Thus we say that the “Q” of the cavity is temporarily spoiled. When most of the atoms are inverted, the cavity Q is restored by switching off the loss. Within a few round trips, the electromagnetic fields in the cavity build up to an extremely high level and almost all of the energy stored in the inverted atomic population is released in a giant pulse. The dynamics of this process are important. The speed with which the entire population can be inverted depends on the pumping rate and the relaxation rates of the pertinent energy levels. This can typically take 10’s to 100’s of round



**Noise and Stability in Modelocked Soliton Lasers, Figure 15**  
**a** Parametric plot of the solutions  $x(t)$ ,  $y(t)$ ,  $z(t)$  of the Lorenz equations for  $r = 160.0$ ,  $\sigma = 1.43$  and  $b = 0.28$ . Now the orbit closes upon itself after one simple traversal. **b**  $x^2(t)$  alone (proportional to power in electric field)

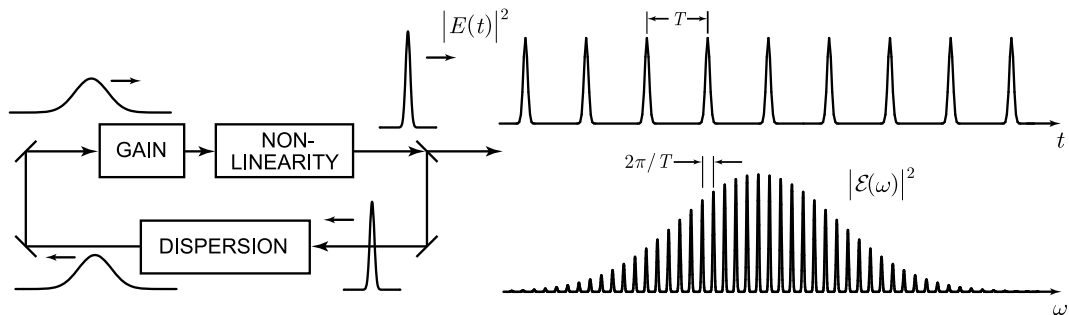
trip cavity-times. The final pulse width is typically on the order of a few to 10's of nanoseconds and the repetition rate can range from  $< 10$  Hz to  $> 10$  KHz.

Another mechanism for producing short pulses with high energy is called “modelocking”. This technique produces far shorter pulses at a much higher repetition rate than Q-switching, albeit at lower energies per pulse. It is the technique of choice for producing the shortest optical pulses, and today, pulses of just a few cycles duration ( $\tau \approx 5$  femtoseconds) in the visible portion of the spectrum are routinely generated in laboratories and soft x-ray pulses have even crossed the 1 femtosecond barrier ( $1 \text{ fs} = 10^{-15} \text{ s}$ ) [51]. The term “modelocked” as applied to lasers arises when one considers the cooperative effect of many thousands, or hundreds of thousands, of spectral comb lines which constitute a train of ultrashort light pulses. Although there are a number of approaches to achieving this phenomenon, we will discuss only the one that is currently responsible for the revolution in ultrashort pulse laser technology and the explosive growth of ultrafast science as well [52].

Consider the elementary laser system shown in Fig. 16. There are four essential components in this model. All oscillators must have gain and feedback, but here we added nonlinearity and dispersion. Let us assume we have a pulse somewhere in the cavity and suppose we follow this pulse around on complete loop. Beginning with the entrance to the gain medium it is easy to understand that the pulse is simply amplified, provided that the gain medium is linear. Next, the pulse enters a nonlinear medium which preferentially passes the highest amplitude portion of the pulse and this causes a narrowing. Such an object is sometimes called a “saturable absorber” but actual energy absorption is not necessary to cause pulse narrowing (refractive effects such as self-focusing produce similar effects). This pulse narrowing is almost always accompanied by self-phase modulation. In this model some of the narrowed

pulse energy escapes the cavity through a partially-transmitting output coupler. Most of the pulse energy stays within the cavity and next passes through a region of dispersion which has the effect of broadening and chirping the pulse. Finally, the broadened pulse is amplified again and we have completed one circuit of the laser. This highly simplified model neglects many subtle, but important, effects such as self-focusing, diffraction losses, bandwidth limiting and dispersion of elements like mirror coatings, etc. However, the basic four elements in this model can create a self-consistent situation that fairly accurately describes the modern generation of “Kerr-lens modelocked soliton lasers”. The Kerr-lensing effect is brought about by the high peak powers in the circulating pulse causing a self-focusing effect in the laser gain medium which is designed to match the other focusing elements in the cavity. This creates a sort of unstable equilibrium in which only a high-power pulse can be a self-consistent solution since the self-focusing is *required* in order to produce a stable cavity as far as the transverse properties of the laser field are concerned. The appellation “soliton” arises from the similarity between the periodic application of nonlinearity and dispersion in the laser to the distributed effects of nonlinearity and dispersion which are required in systems that admit the solitary waves called “solitons” as solutions [53].

The self-consistent picture of a pulse evolving every round-trip within the cavity also implies that the laser produces a train of pulses with spacing,  $T$ , equal to the round-trip cavity time (see Fig. 16). Simple Fourier transformation of such a periodic sequence gives a spectrum of periodically-spaced comb-lines with interline spacing  $\delta\omega = 2\pi/T$  as shown in the figure. The width of the spectrum, and hence the number of comb-lines contained therein, is inversely proportional to the pulsewidth. For example, a 5 fs pulse has a bandwidth  $\Delta\nu \approx 100$  THz which therefore contains about  $10^6$  comb lines if the laser has a 100 MHz repetition rate. The remarkable, and es-



**Noise and Stability in Modelocked Soliton Lasers, Figure 16**  
Principle of operation of the modelocked soliton laser



sential, fact about these spectral lines is that they must all have a well-defined relative phase with respect to each other. Furthermore, these phases must remain constant with time, or be “locked” together. The spectral lines of the pulse train are also, in fact, the allowed eigenmodes or axial modes of the laser cavity and thus we say that the axial modes are locked, or the laser is “modelocked”.

### Solitons

One of the principal features of linear systems is superposition of solutions. That is, if  $u_1(x, t)$  and  $u_2(x, t)$  are separate independent solution to a particular problem in a linear system, then the sum of the two,  $u_1(x, t) + u_2(x, t)$ , is also a solution. One could say that the two solutions are “unaware” of each other’s presence in the linear system. They do not interact or depend upon each other in any way. The same cannot be said for a nonlinear system. The presence of two disturbances in a nonlinear system generally results in interaction between the two or even self-interaction within a single solution.

Except for electromagnetic waves in a vacuum, almost all wave propagation problems are nonlinear in nature. This is because the waves depend upon a medium to transfer their energy. Perturbations about equilibrium of the constituents of the medium eventually drive them into a nonlinear region of response given a large enough driving force and thus the collective wave phenomena must be described by a nonlinear wave equation. An example of such an equation is that which describes low-amplitude shallow water waves. Derived in 1895 by Korteweg and de Vries [54], this has come to be known as the Korteweg–de Vries or KdV equation;

$$\frac{\partial u}{\partial t} - 6u \frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0. \quad (88)$$

This equation was originally developed to describe the long solitary waves that were first observed in a canal by John Scott Russell in 1834 [55].

Assuming a traveling-wave solution of the form  $u(x, t) = u(x - vt)$  then it can be shown that the solution to the KdV equation is [53,56,57]

$$u(x, t) = -\frac{v}{2} \operatorname{sech}^2 \left[ \frac{\sqrt{v}}{2} (x - vt - x_0) \right]. \quad (89)$$

This function represents an isolated pulse traveling with velocity  $v$ .  $x_0$  is the position of the center of the pulse at time  $t = 0$ . Notice that as the velocity increases, so does the amplitude while the width of the pulse decreases. This coupling between the amplitude, velocity and pulsewidth is quite typical of nonlinear waves.

The KdV equation was studied for many years and in 1965, in a seminal paper, Zabusky and Kruskal [58] discovered that the solitary wave solutions (89) to the KdV equation could interact *elastically*. That is, pulses could collide, exchange energy and emerge with the same profile and velocity and only a slight phase shift. This resurrection of pulse form following interaction in a nonlinear medium was unexpected and, owing to the similarity to fundamental particle scattering, the authors coined the name “soliton” for these very special types of solitary waves. The initial paper of Zabusky and Kruskal launched quite a revolution in applied mathematics and soon solitons were predicted and observed in many systems [53,59].

One of the many nonlinear wave equations that admits solitons as solutions is the nonlinear Schrödinger equation (NLS), also called the cubic Schrödinger equation. This is particularly relevant to the world of ultrafast optics since solitons were predicted by Hasegawa and Tappert as solutions to the wave equation in optical fibers under certain conditions [60]. Simply put, the required conditions are that the intrinsic anomalous dispersion of the optical fiber should exactly balance the nonlinearity due to the quadratic electro-optic, or Kerr, effect. This meant that short optical pulses (and hence high data rates) could be transmitted over very long distances without suffering the usual broadening associated with dispersion provided that the peak pulse power was high enough to drive a nonlinear reaction that exactly compensated for the linear dispersion. In an elegant series of experiments, Mollenauer et al., demonstrated this phenomenon and set the stage for all future optical fiber telecommunications technology [61].

The nonlinear Schrödinger equation, in normalized units, is

$$i \frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + 2|u|^2 u = 0. \quad (90)$$

We assume solutions of the general form

$$u(x, t) = \phi(x, t) e^{i\theta(x, t)} \quad (91)$$

where both  $\phi(x, t)$  and  $\theta(x, t)$  are assumed to be real. Substituting these into the NLS equation (90) above, we find that the traveling-wave forms

$$\phi(x, t) = \tilde{\phi}(x - v_e t) \quad (92)$$

$$\theta(x, t) = \tilde{\theta}(x - v_c t) \quad (93)$$

will work as solutions [53], where  $v_e$  and  $v_c$  are the envelope and carrier velocities, respectively. Specifically, we find that

$$\tilde{\phi} = a \operatorname{sech} [a(x - v_e t)] \quad (94)$$

provided that

$$4a^2 = v_e^2 - 2v_e v_c \quad (95)$$

is also satisfied. The final form of the solution can then be shown to be

$$u(x, t) = a \exp \left[ i \frac{v_e}{2} x + i \left( a^2 - \frac{v_e^2}{4} \right) t \right] \text{sech} [a(x - v_e t - x_0)] . \quad (96)$$

Like the solution to the KdV equation, this wave function is also a soliton and variations on it will appear again and again in modelocked laser physics and nonlinear optical fiber analysis.

### Modelocked Soliton Lasers

To see how solitons come into the theory of certain types of modelocked lasers, we develop a self-consistent model of a modelocked laser following a packet of photons around a unidirectional ring cavity (Fig. 17). This will generate the so-called “master-equation” as originally developed by Haus [62] to analyze modelocking with a fast saturable absorber. It was later refined to include nonlinear phase shifts due to self-modulation [63] and the Kerr-lens effect [64] and serves as the starting point for noise and stability analyses [65,66,67].

Consider the schematic of the ring laser shown below (Fig. 17). The physical phenomena related to the passage of a pulse around the ring are grouped into four separate boxes. They account for linear loss and phase shift, gain (or amplification), group velocity dispersion (GVD) and self

phase modulation (SPM) and saturable absorption. The key assumptions in the following analysis are [62] 1) The relaxation time of the saturable absorber is very short compared with the optical pulse width, 2) The saturable absorption is describable using a rate equation approach and the power dependent absorption coefficient is expanded to first order in the power, 3) The relaxation time of the laser gain medium is long compared to the optical pulse width so that the gain is essentially constant, 4) The modelocked pulse is changed by only a small amount upon each circuit of the ring, and 5) The spectrum of the modelocked pulse is small compared with the laser gain bandwidth.

The slowly-varying envelope approximation will allow us to be primarily concerned with the behavior of the envelope of the optical pulse, which we shall denote as  $a(t)$  such that, in suitable units, the electric field is given by  $E(t) = a(t) \exp [i(\omega_0 t - k_0 z)]$ . The total loss due to diffraction, dielectric loss (absorption), and finite mirror reflectivities, as well as the net linear phase shift will produce a complex change in the amplitude

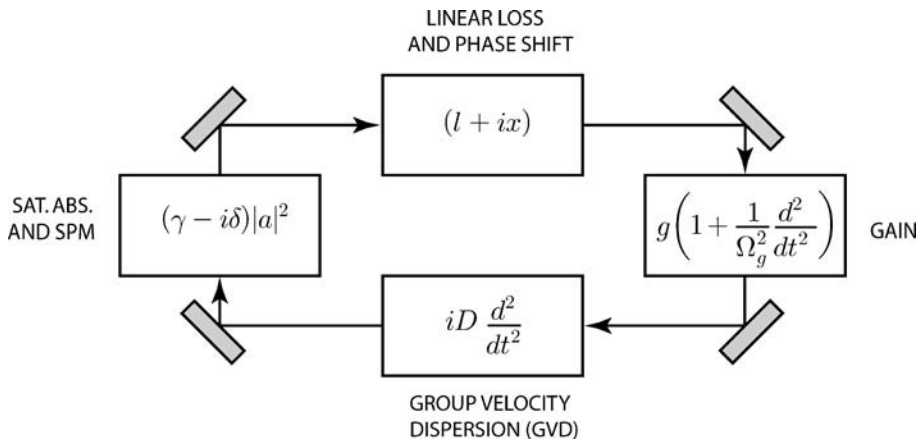
$$\Delta a = -(l + ix)a \quad (97)$$

per pass. For the case of low gain, the change in amplitude per pass is

$$\Delta a = g \left( 1 + \frac{1}{\Omega_g^2} \frac{d^2}{dt^2} \right) a \quad (98)$$

where  $\Omega_g$  is the bandwidth of the gain curve in the parabolic approximation. The group velocity dispersion (GVD) produces a change that is purely imaginary

$$\Delta a = iD \frac{d^2}{dt^2} a . \quad (99)$$



Noise and Stability in Modelocked Soliton Lasers, Figure 17

Unidirectional ring laser cavity for analysis of the modelocked soliton laser

Here  $D$  is the dispersion parameter defined as

$$D = \frac{1}{2} \frac{d^2 k}{d\omega^2} l_{\text{eff}}, \quad (100)$$

where  $k$  is the propagation constant of the net dispersive media in the cavity with an effective length  $l_{\text{eff}}$  (Note: here again we combine all of the dispersion that can occur due to dielectrics, mirror coatings, etc., into one term for convenience).

Self phase modulation (SPM) appears as a term proportional to the peak amplitude

$$\Delta a = -i\delta |a|^2 a \quad (101)$$

where the nonlinear coefficient takes the form

$$\delta = \frac{\omega_0}{c} \frac{n_2 d_{\text{eff}}}{\mathcal{A}_{\text{eff}}}. \quad (102)$$

The nonlinear index  $n_2$  is due to the Kerr effect. It contributes to the index of refraction according to

$$n = n_0 + n_2 I \quad (103)$$

where  $I$  is the intensity ( $W/m^2$ ) of the optical field.  $\mathcal{A}_{\text{eff}}$  is the effective field mode area and  $d_{\text{eff}}$  is the effective length of the nonlinear medium.

We also must include the effect of a fast saturable absorber,

$$\Delta a = \gamma |a|^2 a \quad (104)$$

where  $\gamma$  is the saturation parameter and is inversely proportional to the saturation intensity  $I_{\text{sat}}$ .

Finally, we add a term that allows for small carrier frequency changes away from the cavity mode resonances (Fabry–Perot modes). A shift in frequency of  $\Delta\omega_0$  corresponds to a phase shift per pass of

$$\psi = \frac{\Delta\omega_0}{c} L_{\text{eff}} \quad (105)$$

where  $L_{\text{eff}}$  is the effect round-trip cavity path length. The change in field amplitude is therefore

$$\Delta a = \exp(-i\psi) a - a \approx -i\psi a. \quad (106)$$

We now combine all of these effects by assuming that on a per-pass basis, they are all small and therefore are ad-

ditive. In the steady state, then, they must all sum to zero

$$\left[ -i\psi - (l + ix) + g \left( 1 + \frac{1}{\Omega_g^2} \frac{d^2}{dt^2} \right) + iD \frac{d^2}{dt^2} + (\gamma - i\delta) |a|^2 \right] a = 0. \quad (107)$$

This is the “master equation” of steady-state modelocking [63]. A solution to this equation can be found by introducing the ansatz

$$a(t) = A_0 \text{sech} \left( \frac{t}{\tau} \right) \exp \left[ i\beta \ln \text{sech} \left( \frac{t}{\tau} \right) \right] \quad (108)$$

where  $\beta$  is a so-called “chirp parameter” and  $\tau$  is the pulsewidth. (Note that this form of sech pulse is remarkably similar to the soliton solutions of the KdV and NLS equations, and this version version allows for chirped pulses; that is, pulses with a sweep in frequency). If we substitute the ansatz into the master equation (107), and match coefficients common to the sech and  $\text{sech}^2$  terms, we generate two complex equations in four unknowns,

$$-i\psi + g - l - ix + \frac{(1 + i\beta)^2}{\tau^2} \left( \frac{g}{\Omega_g^2} + iD \right) = 0 \quad (109)$$

$$\frac{2 + 3i\beta - \beta^2}{\tau^2} \left( \frac{g}{\Omega_g^2} + iD \right) = (\gamma - i\delta) A_0^2. \quad (110)$$

The four unknowns determined by these equations are the pulsewidth,  $\tau$ , the chirp parameter,  $\beta$ , the phase shift,  $\psi$ , and the gain,  $g$ . The pulse amplitude  $A$  is assumed to be such that the gain saturates down to a level that is approximately equal to the loss ( $g \approx l$ ) where

$$g = \frac{g_0}{1 + 2A^2\tau/P_s T_R} = \frac{g_0}{1 + W/P_s T_R} \quad (111)$$

where the pulse energy is given by  $W = 2A^2\tau$ ,  $P_s$  is the effective saturation power and  $T_R$  is the round-trip cavity time.

The complete solutions for  $\tau$ ,  $\beta$ ,  $\psi$  and  $g$  and the conditions for their realization need not be elaborated here since they are covered in detail elsewhere [63,64]. Suffice it to say that steady-state solutions of a soliton nature do exist and form the basis for most modern ultrashort-pulse solid-state lasers [52]. Pulses of  $< 10$  fs duration with peak powers of several hundred kilowatts and repetition rates of 50–1000 MHz are now routinely available in laboratories worldwide.

We next move on to learn about the fundamental processes that govern and generate noise in both CW and modelocked lasers and finally we will return the specific

case of the modelocked soliton laser to see how the important parameters, derived above, vary when subject to perturbations, for this is at the heart of the issue of instability.

### Laser Noise and Linewidth

Noise accompanies all signals and from an elementary Fourier transform point of view, noise necessarily broadens the spectral linewidth of any source. One can appreciate this causal relationship better using the concepts of modulation which will be developed in Subject. “[Analytical Description of Envelope Noise](#)”. We can distinguish between two sources of noise in a laser field. The first, sometimes called “technical noise” arises from environmental perturbations to the oscillator which creates the field. These are controllable to a greater or lesser degree and are not fundamental in the sense that they are not necessary parts of the amplification or oscillation process. The second source of noise is the spontaneous emission of photons and arises due to the large number of inverted atoms (or molecules) required to produce gain through stimulated emission. Spontaneous emission occurs with equal probability in all directions but some of these photons will radiate directly into the desired laser cavity mode.

Since the generation of the spontaneously emitted photons is a random process, they are not coherent with the desired mode and thus constitute noise. If it were not for the presence of these noise photons, the laser electromagnetic field would be perfectly coherent and measurement of the spectrum of the field with an ideal spectrometer would yield a single narrow line at the oscillation frequency. The width of this line,  $\Delta\nu$ , would be determined solely by the inverse of the time spent in its measurement:  $\Delta\nu \approx 1/\Delta t_{\text{meas}}$ . However, the addition of spontaneous emission photons to the ideal coherent field produces a small perturbation to the amplitude and phase of the field which, in turn, will broaden the width of the spectrum. This noise mechanism sets a fundamental quan-

tum limit to the minimum achievable laser linewidth and is therefore an essential concept in laser physics. We now proceed to calculate the nature of this noise and how it affects the linewidth of the desired field. (This treatment follows that of Yariv [10].)

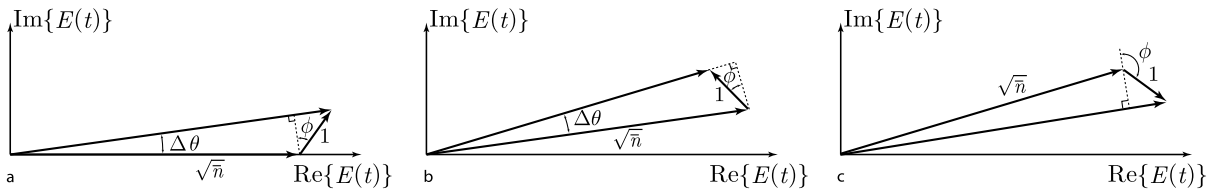
### Oscillator Spontaneous Emission Noise

In order to incorporate the presence of spontaneous emission into the coherent field of a laser oscillator, we write the field as

$$\mathcal{E}(t) = \text{Re} \left[ E(t) e^{i(\omega_0 t + \theta(t))} \right] \quad (112)$$

where  $E(t)$  and  $\theta(t)$  are random processes whose fluctuations are very small about their mean values. Spontaneous emission noise can be modeled as adding to the otherwise perfectly coherent field a photon at the frequency of the transition,  $(E_2 - E_1)/h$ . The *phasor picture* of the field with the addition of this randomly oriented spontaneous photon provides a most satisfactory view of the process (see Fig. 18). We will refer to the unperturbed electromagnetic field as the *carrier*, the magnitude of which is proportional to the square root of the mean number of photons in the field mode;  $|E| \propto \sqrt{\bar{n}}$ . The addition of one spontaneous emission photon to the carrier is accomplished by adding to the electric field phasor a unit-amplitude phasor ( $n = 1$ ) with random phase which we will call  $\phi$  and measure it arbitrarily with respect to the normal to the carrier phasor following the addition of the noise photon. The phase angle  $\phi$  is uniformly distributed on the interval  $[0, 2\pi]$  and thus the probability density function  $p(\phi) = 1/2\pi$ . In general, the number of carrier photons will be quite large ( $\bar{n} \gg 1$ ) and thus the phase deviation of the carrier phasor due to a single spontaneous emission photon,  $\Delta\theta(1)$ , can be found from the projection

$$1 \cdot \cos \phi = \sqrt{\bar{n}} \sin \Delta\theta(1) \approx \sqrt{\bar{n}} \Delta\theta(1) \quad (113)$$



**Noise and Stability in Modelocked Soliton Lasers, Figure 18**

Phasor picture of the addition of a spontaneous emission photon (amplitude = 1) to the electric field optical carrier  $E(t)$  (amplitude =  $\sqrt{\bar{n}}$  where  $\bar{n}$  is the mean photon number in the carrier field). The phase angle  $\phi$  of the spontaneous emission photon is a random variable and thus its projection on the carrier phasor,  $\cos \phi$ , introduces phase ( $\Delta\theta$ ) and amplitude perturbations to the electric field. The sequence a–c depicts three spontaneous emission events adding to the carrier field. Since the mean photon number is typically very large,  $\bar{n} \gg 1$ , the phasor lengths here are highly exaggerated

or

$$\Delta\theta(1) = \frac{1}{\sqrt{\tilde{n}}} \cos \phi. \quad (114)$$

The carrier phase deviation  $\Delta\theta$  is a random variable and its statistics are what determine the lineshape. Since  $\phi$  is uniformly distributed on the interval  $[0, 2\pi]$ , the mean value of the carrier phase deviation after a large number  $N_{\text{sp}}$  of spontaneous emission events is obviously zero. This can be shown simply;

$$\begin{aligned} \langle \Delta\theta(N_{\text{sp}}) \rangle &= \int_0^{2\pi} \Delta\theta(1) p(\phi) d\phi \\ &= \frac{1}{2\pi\sqrt{\tilde{n}}} \int_0^{2\pi} \cos \phi d\phi = \frac{1}{2\pi\sqrt{\tilde{n}}} \cdot 0 = 0. \end{aligned} \quad (115)$$

The *mean-squared* value of the phase deviation, on the other hand, is non-zero and is given from the results of the random walk process [68],

$$\langle [\Delta\theta(N_{\text{sp}})]^2 \rangle = \langle [\Delta\theta(1)]^2 \rangle N_{\text{sp}} = \frac{1}{\tilde{n}} \langle \cos^2 \phi \rangle N_{\text{sp}} = \frac{N_{\text{sp}}}{2\tilde{n}}. \quad (116)$$

This is the mean-squared phase deviation after  $N_{\text{sp}}$  spontaneous emission photons have been emitted and combined with the carrier field. The next task is to convert this expression to an r.m.s. phase deviation in a given time interval  $\tau$ . We must first calculate the average number of spontaneous emission events in a time  $\tau$ . The total number of spontaneously emitted photons per second into *all modes* (i. e. the spontaneous emission rate) is given by

$$\tilde{N}_{\text{sp}}(s^{-1}) = \frac{N_2}{\tau_{\text{rad}}} = N_2 \gamma_{\text{rad}}. \quad (117)$$

The tilde on  $N_{\text{sp}}$  serves to indicate that this is a *rate* as opposed to a simple number. Since all of the spontaneous photons are emitted into all possible modes ( $4\pi$  steradians), we need to normalize this to the total number of modes,  $p$ , within the atomic linewidth

$$\tilde{N}_{\text{sp}}(s^{-1} \text{mode}^{-1}) = \frac{N_2}{\tau_{\text{rad}} p} = N_2 \frac{\gamma_{\text{rad}}}{p}. \quad (118)$$

Now, recall that the population inversion at threshold is given by

$$\Delta N_{\text{th}} \equiv (N_2 - N_1)_{\text{th}} = \frac{\gamma_c}{\gamma_{\text{rad}}} p \quad (119)$$

where  $\gamma_c$  is the energy decay rate of the cavity. We can now use this result to eliminate the mode number  $p$  from (118),

$$\begin{aligned} \tilde{N}_{\text{sp}} &= N_2 \frac{\gamma_{\text{rad}}}{p} \left( \frac{\Delta N_{\text{th}}}{\Delta N_{\text{th}}} \right) = \frac{N_2}{\Delta N_{\text{th}}} \frac{\gamma_{\text{rad}}}{p} \Delta N_{\text{th}} \\ &= \frac{N_2}{\Delta N_{\text{th}}} \frac{\gamma_{\text{rad}}}{p} \frac{\gamma_c}{\gamma_{\text{rad}}} p \end{aligned} \quad (120)$$

$$= \frac{N_2}{\Delta N_{\text{th}}} \gamma_c = \frac{N_2}{\Delta N_{\text{th}}} \frac{1}{\tau_c}. \quad (121)$$

This is the spontaneous emission rate into a single mode. The total number of spontaneous photons into this mode in a time  $\tau$  is therefore

$$N_{\text{sp}}(\tau) = \tilde{N}_{\text{sp}} \tau = \frac{N_2}{\Delta N_{\text{th}}} \frac{\tau}{\tau_c}. \quad (122)$$

We can now use this result in (116) to obtain the mean-squared phase deviation in a time  $\tau$ ,

$$\langle [\Delta\theta(N_{\text{sp}}(\tau))]^2 \rangle = \frac{\tau}{2\tilde{n}\tau_c} \frac{N_2}{\Delta N_{\text{th}}} \quad (123)$$

and thus the r.m.s. phase deviation is

$$\Delta\theta_{\text{rms}}(\tau) \equiv \sqrt{\langle [\Delta\theta(N_{\text{sp}}(\tau))]^2 \rangle} = \sqrt{\frac{\tau}{2\tilde{n}\tau_c} \frac{N_2}{\Delta N_{\text{th}}}}. \quad (124)$$

Because the amplitude and phase of the electric field (112) incorporate random processes, the Fourier transform does not exist, but the power spectrum  $S_{\mathcal{E}}(\omega)$  does and is related to the Fourier transform of the autocorrelation function of the electric field  $R_{\mathcal{E}}(\tau)$  as established by the Wiener-Khinchine theorem [69]

$$S_{\mathcal{E}}(\omega) = \int_{-\infty}^{\infty} R_{\mathcal{E}}(\tau) e^{-i\omega\tau} d\tau = \mathcal{F}\{R_{\mathcal{E}}(\tau)\}. \quad (125)$$

The autocorrelation function is defined by

$$R_{\mathcal{E}}(\tau) \equiv \langle \mathcal{E}(t) \mathcal{E}(t+\tau) \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T \mathcal{E}(t) \mathcal{E}(t+\tau) dt. \quad (126)$$

Now, because the autocorrelation function contains expectation values of the electric field, we need to provide probability density functions for the random variables. At this point, we note that for an oscillator, the gain is, by definition, saturated, and thus there is not much amplitude fluctuation so we let  $E(t) = E$ . The majority of the spontaneous emission contributes to phase fluctuations. Since  $\Delta\theta$  is the result of a large number of statistically independent events, we can apply the central limit theorem which



says that all such independent random processes approach a Gaussian distribution in the limit of a large number of samples [68,69]. Thus, a normalized Gaussian distribution with mean-squared value  $\langle(\Delta\theta)^2\rangle$  is given by

$$p(\Delta\theta) = \frac{1}{\sqrt{2\pi \langle(\Delta\theta)^2\rangle}} e^{-(\Delta\theta)^2/2\langle(\Delta\theta)^2\rangle}. \quad (127)$$

Since we have already calculated the mean-squared value of the phase fluctuation for the spontaneous emission (123) the autocorrelation function can be found and then we can proceed to calculating the power spectral density (the details of which will be omitted for brevity but can be found in [10])

$$S_{\mathcal{E}}(\omega) = \int_{-\infty}^{\infty} R_{\mathcal{E}}(\tau) e^{-i\omega\tau} d\tau \quad (128)$$

$$= \frac{\langle E^2 \rangle}{2} \frac{N_2}{4\hbar\tau_c\Delta N_{\text{th}}} \left[ \frac{1}{\left(\frac{N_2}{4\hbar\tau_c\Delta N_{\text{th}}}\right)^2 + (\omega + \omega_0)^2} + \frac{1}{\left(\frac{N_2}{4\hbar\tau_c\Delta N_{\text{th}}}\right)^2 + (\omega - \omega_0)^2} \right]. \quad (129)$$

We see that the spectral density is composed of two Lorentzian lineshapes at  $\omega = \pm\omega_0$ . The frequencies for which this Lorentzian is reduced to half of its value on resonance are

$$\omega_{\pm 1/2} = \omega_0 \pm \frac{N_2}{4\hbar\tau_c\Delta N_{\text{th}}}. \quad (130)$$

The full linewidth at half-maximum (FWHM) is therefore

$$\Delta\omega_{\text{osc}} = \omega_{1/2} - \omega_{-1/2} = \frac{N_2}{2\hbar\tau_c\Delta N_{\text{th}}} \quad (131)$$

or, from (123), we have

$$\Delta\omega_{\text{osc}} = \frac{\langle[\Delta\theta(N_{\text{sp}}(\tau))]^2\rangle}{\tau}. \quad (132)$$

The Lorentzian linewidth is just the mean-squared phase deviation per unit time interval  $\tau$ .

We can rewrite (131) in a slightly different, but useful, form by noting that the total power emitted by the atoms into the oscillating mode is  $P_{\text{osc}} = \hbar\omega_0/\tau_c$ . We also note that the cold-cavity linewidth (FWHM) of the resonator is given by  $\Delta\nu_c = (2\pi\tau_c)^{-1}$ . Using these in (131) above we have

$$\begin{aligned} \Delta\omega_{\text{osc}} &= \frac{\hbar\omega_0 N_2}{2P_{\text{osc}}\tau_c^2\Delta N_{\text{th}}} = \frac{\hbar\omega_0 N_2 (2\pi\Delta\nu_c)^2}{2P_{\text{osc}}\Delta N_{\text{th}}} \\ &= \frac{2\pi^2\hbar\omega_0 N_2 \Delta\nu_c^2}{P_{\text{osc}}\Delta N_{\text{th}}} \end{aligned} \quad (133)$$

or

$$\Delta\nu_{\text{osc}} = \frac{\Delta\omega_{\text{osc}}}{2\pi} = \frac{\pi\hbar\nu_0}{P_{\text{osc}}} \frac{N_2}{\Delta N_{\text{th}}} \Delta\nu_c^2. \quad (134)$$

This fundamental result is known as the ‘‘Schawlow–Townes limit’’ for the linewidth of a laser oscillator [70]. Due to many other processes (described previously as ‘‘technical noise’’), this ultimate linewidth is rarely achieved in practice but serves as an important benchmark for precision spectroscopy and metrology [8,9]. To get a feeling for the narrowness of the limit imposed by (134) let’s assume a helium-neon laser ( $\lambda = 632.8$  nm) with a 30 cm mirror spacing and an output coupler reflectivity of  $R = 98\%$ . We further assume that the lower energy level is always empty so that  $\Delta N_{\text{th}} \approx N_2$ . Thus, the cold cavity linewidth is

$$\Delta\nu_c = \frac{1-R}{2\pi} \Delta\nu_{\text{ax}} = \frac{(1-R)c}{4\pi L} \approx 1.6 \text{ MHz} \quad (135)$$

and for 1 mW output power, the laser linewidth is

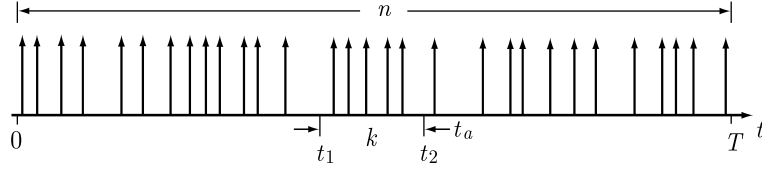
$$\Delta\nu_{\text{osc}} = \pi \frac{\hbar c \Delta\nu_c^2}{\lambda_0 P_{\text{osc}}} = 2.5 \times 10^{-3} \text{ Hz}. \quad (136)$$

This theoretical linewidth is always overwhelmed by acoustic, thermal and other perturbations on the laser cavity and typically the actual measured linewidth is of order  $10^4$ – $10^6$  Hz [9].

When lasers are used to probe narrow atomic or molecular transitions, it is often the laser linewidth that sets the ultimate resolution in the frequency domain. In practice, the transitions of greatest interest often challenge the performance of the probing lasers and this has been responsible for many advances in narrow-linewidth lasers. For example, by stabilizing the cavity length of a laser against a separate high-Q reference cavity, linewidths at the  $\approx 1$  Hz level have been achieved [71,72,73].

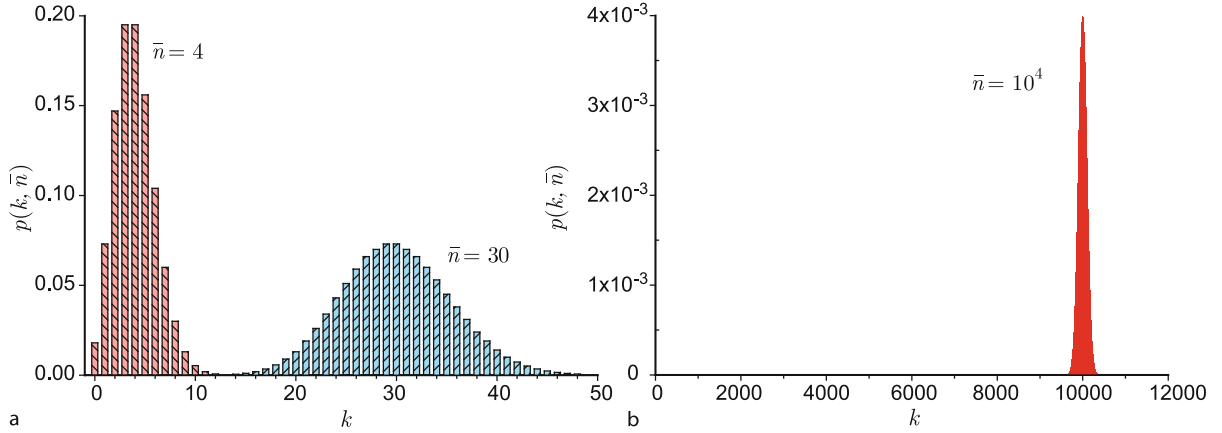
### Shot Noise

The calculation of linewidth in the preceding section gives an idea of the limitations of a laser field when considering its use as a frequency reference or measurement tool. There is another type of uncertainty that enters into optical metrology that occurs when we actually measure the power (or intensity; power-per-unit area) in the field. When measuring the power an optical field we are, in effect, counting photons. This is because optical detectors rely on the conversion of field energy into quantized charge units (e. g. electrons). The photogeneration of electrons is therefore a necessarily quantized process since energy is removed from the radiation field in quantized units



**Noise and Stability in Modelocked Soliton Lasers, Figure 19**

Random sequence of photoelectrons generated in a detector exposed to a laser field. If it is known that  $n$  events are randomly distributed over the interval  $T$ , the probability of finding  $k$  events in the time interval  $t_a = t_2 - t_1$  is given by the Poisson distribution. Note that the average rate of events is  $\bar{N} \equiv n/T$



**Noise and Stability in Modelocked Soliton Lasers, Figure 20**

Poisson probability distributions for several cases of mean photon numbers. **a** Left;  $\bar{n} = 4$  and Right;  $\bar{n} = 30$ . Notice the tendency toward Gaussian shape as  $\bar{n}$  increases. **b** For even higher values of  $\bar{n}$ , (e.g.  $\bar{n} = 10^4$ ) width of the probability distribution becomes a decreasing fraction of the mean value indicating an improvement in signal-to-noise ratio with  $\bar{n}$

sufficient to liberate the charge carrier. Furthermore, the generation of a charge carrier is completely uncorrelated with the generation of any previous charge carriers. That is, the events occur independently of each other. Figure 19 depicts the generation of electrons due to photons hitting a detector. If we make many measurements during a time interval  $t_a = t_2 - t_1$  we will find, on average,  $\bar{n}$  photoelectrons but each individual measurement will yield a number  $k$  that will deviate from  $\bar{n}$  by an amount given by the statistics of the process. The statistics of the generation of photoelectrons is that of a Poisson process [68,74,75] and we will derive several features of it now.

Suppose that  $n$  independent random events occur in a time interval  $T$ , as shown in Fig. 19. We are interested in finding the probability that  $k$  such events occur in a subinterval  $t_a = t_2 - t_1$ . This is a problem in Bernoulli trials and combinatorics [68,69] with the result that

$$\begin{aligned} p_n(k) &= \text{Probability}\{k \text{ events in interval } (t_1, t_2)\} \\ &= \frac{n!}{k!(n-k)!} p^k q^{n-k} = \binom{n}{k} p^k q^{n-k} \end{aligned} \quad (137)$$

where

$$p \equiv \frac{t_2 - t_1}{T} \quad (138)$$

is the probability of a single event in the interval  $t_2 - t_1$  if only one event occurs during the longer time  $T$ , and  $q = 1 - p$ .

The value of  $k$  that gives the maximum probability of  $p_n(k)$  is called the “most likely” value of  $k$  and is found to be about  $np$  [68]. Noting that  $n$  events occur in time  $T$ , we remark that the average rate of occurrences in time  $T$  is  $n/T \equiv \bar{N}$  events/second and therefore we may interpret

$$np = \frac{n}{T}(t_2 - t_1) = \bar{N}(t_2 - t_1) \equiv \bar{n} \quad (139)$$

i. e. the *average* number of events in the interval  $t_2 - t_1$ .

**Asymptotic Forms: Limit of Few Events** If the most likely value of  $k = np$  is of order  $np \approx 1$ , then a useful approximation to (137) is given by Poisson’s theorem [76]

$$\binom{n}{k} p^k q^{n-k} \simeq e^{-np} \frac{(np)^k}{k!} \quad (140)$$

and since we have identified  $np$  as the average number of events in the interval  $(t_2 - t_1)$ , we may write this in the simpler form

$$p(k, \bar{n}) = e^{-\bar{n}} \frac{\bar{n}^k}{k!}. \quad (141)$$

This is the *Poisson Probability Distribution* and is shown in Fig. 20 for several interesting cases. In Fig. 20a, we see two sets of distributions for mean photon numbers  $\bar{n} = 4$  and  $\bar{n} = 30$ , respectively. Clearly as the mean photon number increases, the width of the distribution also increases while the peak probability is reduced. This is consistent with the notion that the sum of probabilities over all outcomes must be unity;

$$\sum_{k=0}^{\infty} e^{-\bar{n}} \frac{\bar{n}^k}{k!} = e^{-\bar{n}} \sum_{k=0}^{\infty} \frac{\bar{n}^k}{k!} = e^{-\bar{n}} e^{\bar{n}} = 1 \quad (142)$$

or, physically, the probability that  $k$  will take on some value between 0 and  $\infty$  is 1.

What is not quite as obvious from the two distributions in Fig. 20a is that the width of the distribution *relative to its mean* is going down as the mean increases. This is shown dramatically in Fig. 20b for the case of  $\bar{n} = 10^4$ . Physically, as we detect a larger number of photons, the relative uncertainty is decreasing. Or, the signal-to-noise ratio improves with the average flux of photons. We will investigate this in a little more detail shortly.

One final remark about the Poisson distribution is the apparent approach to a Gaussian shape as the mean photon number increases. This is not merely a trick of the eye but has its roots in a theorem quoted in the next section.

**Asymptotic Forms: Limit of Many Events** If the total number of events  $n$  in the interval  $T$  is large and so too is the target interval  $(t_2 - t_1)$  such that  $npq \gg 1$ , then in the vicinity of the most likely value of the distribution where  $|k - np| \leq \sqrt{npq}$  the theorem of DeMoivre and Laplace shows that the following approximation is valid [76]

$$p_n(k) = \binom{n}{k} p^k q^{n-k} \simeq \frac{1}{\sqrt{2\pi npq}} e^{-(k-np)^2/2npq}. \quad (143)$$

That is, the Poisson distribution tends toward a Gaussian with mean value  $np$  and standard deviation  $\sigma = \sqrt{npq}$ .

**The Moments** The important statistical properties of a random process are given by their moments. The  $n$ th moment of the random variable  $k$  is defined as the expectation value of  $k^n$ ; that is  $E\{k^n\}$ . The first moment is the “mean” and the second moment is the “mean-square” and

**Noise and Stability in Modelocked Soliton Lasers, Table 4**

**Important moments for the Poisson probability distribution function  $p(k, \bar{n})$**

PROBABILITY	MEAN	MEAN-SQUARE	VARIANCE	STD. DEVIATION
$p(k, \bar{n})$	$\bar{k}$	$\bar{k}^2$	$\sigma^2 = \overline{(k - \bar{n})^2}$	$\sigma = \sqrt{\overline{(k - \bar{n})^2}}$
$= e^{-\bar{n}} \frac{\bar{n}^k}{k!}$	$\bar{n}$	$\bar{n}^2 + \bar{n}$	$\bar{n}$	$\sqrt{\bar{n}}$

so on. The central moments are similar except that we subtract the mean value before finding the expectation. The  $n$ th central moment is given by  $E\{(k - \bar{k})^n\}$  and clearly equals zero for  $n = 1$  but the case  $n = 2$  is so important and useful it is given the special name “variance”. In effect, the variance  $\sigma^2 \equiv E\{(k - \bar{k})^2\}$  is a measure of the mean-square fluctuation of the random variable and its square root,  $\sigma$ , is called the “standard deviation”. These important quantities are straightforward to calculate and are tabulated in Table 4.

From the moments we can immediately say something about measurement uncertainty. If we make a large number of measurements of the number of photons  $k$  in an interval  $t_a$ , then the standard deviation is a measure of the width of the distribution of the measured values. The ratio of this width to the average is a measure of the uncertainty in the measurement

$$\text{Uncertainty in } k = \frac{\sigma}{\bar{k}} = \frac{\sqrt{\bar{n}}}{\bar{n}} = \frac{1}{\sqrt{\bar{n}}}. \quad (144)$$

The uncertainty in the measurement improves (is reduced) as the square root of the average number of detected photons.

**The Power Spectrum** Since the photodetection process produces events that are randomly distributed in time, it does not possess a valid transform. However, a photodetector will produce a physical current when driven by the process and thus power will be transferred to a load. There are many cases when a small, coherent signal will be embedded in the optical field which will have a narrow spectral feature which can be discriminated against the random noise background. It thus is reasonable to search for an expression that describes the spectral distribution of the power, since there are fluctuations, and this is what is contained in the *power spectral density* function, which we calculate now.

The Wiener-Khinchine theorem [69] states that the power spectrum of a random process is equal to the Fourier transform of its autocorrelation. Thus, we need to

find the autocorrelation function of the process describing the photon detection. If  $t_i$  are the random points in time when a photon is detected, then the process of detected photons is described by

$$z(t) = \sum_j \delta(t - t_j) \quad (145)$$

where  $\delta(t)$  is the Dirac delta function defined so that

$$\int_{-\infty}^{\infty} \delta(t) dt = 1 \quad (146)$$

and the points  $t_j$  follow the Poisson distribution (141). The autocorrelation function for this process is [68]

$$R_z(\tau) = E \{z(t)z(t + \tau)\} = \bar{N}^2 + \bar{N} \delta(\tau). \quad (147)$$

We now take the Fourier transform to find the power spectrum

$$S_z(\nu) = \mathcal{F} \{R_z(\tau)\} \quad (148)$$

$$= \int_{-\infty}^{\infty} R_z(\tau) e^{-i2\pi\nu\tau} d\tau \quad (149)$$

$$= \int_{-\infty}^{\infty} [\bar{N}^2 + \bar{N} \delta(\tau)] e^{-i2\pi\nu\tau} d\tau \quad (150)$$

$$= \bar{N}^2 \delta(\nu) + \bar{N} \quad (-\infty \leq \omega \leq \infty). \quad (151)$$

In the last line we emphasize that this is a *double-sided* power spectrum. Therefore, to find the total power passing through a rectangular filter of bandwidth  $B$ , centered at  $\nu_o$ , away from  $\nu = 0$ , we would multiply the spectral density by 2. That is,  $P(\nu_o) = 2S_z(\nu_o)B = 2B\bar{N}$ .

We remark here about a subtle issue regarding the frequency domain variable  $\omega \equiv 2\pi\nu$  with units of radians/second. Almost all instrumentation and discussions of frequency-domain phenomena use the more common variable  $\nu$  with units of Hertz or, the historical predecessor “cycles/second”. In any case, we can define a spectral density function using either variable but the integrals over the relevant domains must yield the same power. Therefore, since  $d\omega = 2\pi d\nu$ , we have

$$\int_{-\infty}^{\infty} S(\omega) d\omega = \int_{-\infty}^{\infty} S(\omega) 2\pi d\nu = \int_{-\infty}^{\infty} S(\nu) d\nu \quad (152)$$

and thus

$$S(\nu) = 2\pi S(\omega). \quad (153)$$

**Photocurrent** Suppose the photodetection process yields a single electron for every incident photon. Then we can describe the current from the photodetector in terms of the random process (145) as

$$i(t) = q \sum_j \delta(t - t_j). \quad (154)$$

The average value of the current measured over a time interval  $T$  is simply

$$\overline{i(t)} = \frac{q}{T} \int_{-T/2}^{T/2} \delta(t - t_j) dt = q \frac{\bar{n}}{T} = q\bar{N} \equiv i_{\text{avg}} \quad (155)$$

since, by definition, there are on average,  $\bar{n}$  events detected in a time interval  $T$ .

Using our previous results for the autocorrelation function and the power spectrum of the random process described by Poisson statistics, we have for the photocurrent

$$R_i(\tau) = E \{i(t)i(t + \tau)\} = q^2 [\bar{N}^2 + \bar{N} \delta(\tau)] \quad (156)$$

and

$$|I(\nu)|^2 = \mathcal{F} \{R_i(\tau)\} = q^2 [\bar{N}^2 \delta(\nu) + \bar{N}] \quad \text{A}^2/\text{Hz} \quad (157)$$

valid for  $(-\infty \leq \nu \leq \infty)$ . Thus, we see that the power spectrum is comprised of two principal features. The delta function at DC ( $\nu = 0$ ) is the power due to the average photocurrent. The second term, with uniform (two-sided) spectral spectral density, describes the power density due to the random process, or the noise. If the photocurrent drives a load resistor of value  $R_L$ , we can identify the average (or DC) power and the noise power within a bandwidth  $B$  as

$$P_{\text{DC}} = q^2 \bar{N}^2 R_L = i_{\text{avg}}^2 R_L \quad \text{W} \quad (158)$$

$$P_{\text{NOISE}} = 2q^2 \bar{N} B R_L = 2q i_{\text{avg}} B R_L \quad \text{W}. \quad (159)$$

Now we have expressions from which we can determine a signal-to-noise ratio for the detected photoelectrons. If we consider the DC photocurrent to be the desired measured quantity, then the signal-to-noise ratio (SNR) for shot-noise-limited detection is given by

$$\text{SNR} = \frac{P_{\text{DC}}}{P_{\text{NOISE}}} = \frac{i_{\text{avg}}}{2qB} \quad (160)$$

where  $B$  is the bandwidth of the measuring device. If the desired signal is a modulation of the average photocurrent,

then it will show up in the spectrum as a coherent spike with power proportional to  $P_{DC}$  and the SNR will scale according to this proportionality. The bandwidth  $B$  is still that of the measuring instrument and often it is a narrow-band filter centered at the frequency of the modulation in order to optimize the SNR.

Since it is often the case that the load resistance and bandwidth of the following circuitry are flexible parameters, it is convenient to note that the single-sided ( $\nu \geq 0$ ) mean-squared noise current spectral density is simply

$$I_{NOISE}(\nu) \equiv 2qi_{avg} A^2/\text{Hz}. \quad (161)$$

Then, for purposes of circuit modeling and analysis, we can include a signal generator with rms current equal to  $i_{rms} = \sqrt{2qi_{avg}} A/\sqrt{\text{Hz}}$ .

### Analytical Description of Envelope Noise

The noise that we will be considering in this section is that pertaining to the amplitude and phase of the laser *intensity envelope*. Instability of the optical carrier as described in Sect. “Laser Noise and Linewidth” is, indeed, an important phenomenon but it is assumed that the envelope of the electric field (and thus the intensity) carries much greater relevance in the case of modelocked lasers since we normally detect power. Furthermore, modelocking imparts a certain degree of phase coherence, or cooperation, among the many longitudinal cavity modes that make up an ultrashort pulse and it becomes rather nebulous as to what is meant by “carrier” amplitude and phase noise. So, we focus our attention on the stability of the modelocked pulse train; both amplitude and phase.

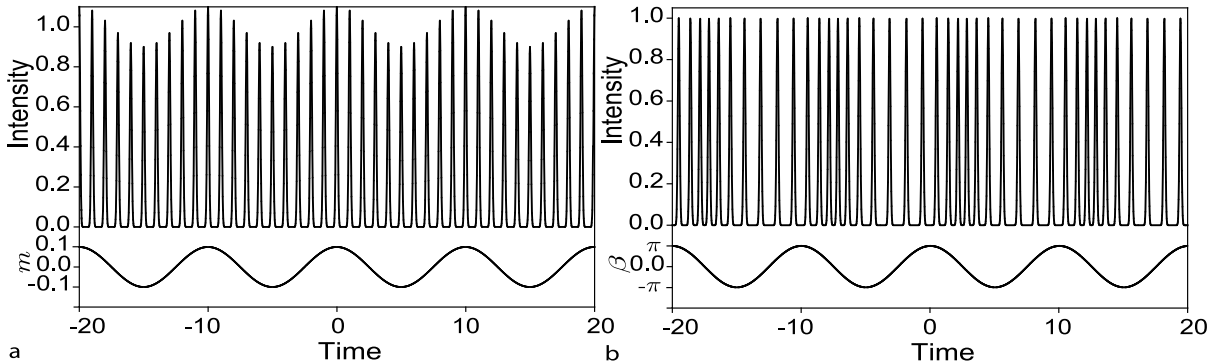
Suppose we model the laser pulse train as a periodic sequence of Gaussian pulses

$$I(t) = \sum_n I_o \exp \left[ -4 \ln 2 \left( \frac{t - nT}{\tau} \right)^2 \right] \quad (162)$$

where  $I_o$  is the peak power,  $T$  is the period and the scaling factor  $4 \ln 2$  ensures that the pulsewidth  $\tau$  represents the full-width at half-maximum (FWHM) of the pulse intensity. The form shown in (162) represents an ideal train of pulses, free of amplitude or timing fluctuations. In order to see the effects of random fluctuations in pulse amplitude and period, we take the point of view that any wideband noise process can be decomposed into narrowband slices which are quasi-coherent on a time scale that is roughly the inverse of the bandwidth of the spectral slice under study [15]. We can then introduce these effects into the model as pure sinusoids by writing

$$I(t) = \sum_n I_o (1 + m \cos \omega_{am} t) \cdot \exp \left[ -4 \ln 2 \left( \frac{t - [n + (\beta/2\pi) \cos(\omega_{pm} t)]T}{\tau} \right)^2 \right]. \quad (163)$$

Here we see that the period is being modulated at a rate  $\omega_{pm}$  and the amplitude at a rate  $\omega_{am}$ . The coefficients  $m$  and  $\beta$  are the amplitude- and phase-modulation indices, respectively.  $\beta$  is also frequently referred to as the “peak phase deviation”. When  $\beta = 2\pi$  the pulse train has undergone one complete period of timing modulation while  $m = 1$  causes 100% amplitude modulation. These two effects are shown graphically in Fig. 21. For purposes of il-



**Noise and Stability in Modelocked Soliton Lasers, Figure 21**

**a** Train of optical pulses from a modelocked laser undergoing pure sinusoidal AM modulation with a modulation index of 10% ( $m = 0.1$ ). **b** Optical pulse train undergoing pure PM modulation with a peak phase deviation of  $\pi$  radians ( $\beta = \pi$ ). Modulation waveform shown in bottom of each figure



illustration, values of  $m = 0.1$  and  $\beta = \pi$  were chosen and the modulation frequencies are 10% of the pulse repetition rate. For real sources, typically  $m \ll 1$  and  $\beta \ll 2\pi$ . Also, the modulation rates  $\omega_{am}$  and  $\omega_{pm}$  are almost always much less than the laser pulse repetition rate. That is,  $\omega_{am}, \omega_{pm} \ll 2\pi/T$ .

### Fourier Analysis of Sinusoidal Amplitude Modulation

The effects of amplitude and phase modulation on an infinite pulse train produce markedly different effects, yet, it turns out that if we simply measure the power spectrum associated with the two cases, the spectra will often be indistinguishable, especially for small modulation depths. To see why this is so and to understand the necessity for experimental tools that can differentiate between them, we analyze the modulation effects on each of the Fourier components in (163) using the phasor representation. Consider the effect of AM modulation first. We expand the sequence of Gaussian pulses in a Fourier series assuming that the pulsewidth is much shorter than the period,  $\tau \ll T$  (as is always the case for modern modelocked lasers)

$$\begin{aligned} I_{AM}(t) &= I_o (1 + m \cos \omega_{am} t) \\ &\cdot \sum_n \exp \left[ -4 \ln 2 \left( \frac{t - nT}{\tau} \right)^2 \right] \quad (164) \\ &= I_o (1 + m \cos \omega_{am} t) \\ &\cdot \frac{\tau}{T} \sqrt{\frac{\pi}{4 \ln 2}} \left[ \frac{1}{2} + \sum_n \cos(2\pi nt/T) \right] \quad (165) \\ &= I'_o \left\{ \frac{1}{2} + \frac{m}{2} \cos(2\pi f_{am} t) + \sum_{n=1}^{\infty} \cos(2\pi nt/T) \right. \\ &\quad \left. + \frac{m}{2} \left[ \cos 2\pi \left( \frac{n}{T} - f_{am} \right) + \cos 2\pi \left( \frac{n}{T} + f_{am} \right) \right] \right\}. \quad (166) \end{aligned}$$

This form illuminates the important components produced by amplitude modulation. The first term in (166) is the average power carried by the pulse train and the second term is the modulation of that average power. The last three terms comprising the infinite series are harmonics of the fundamental repetition rate ( $n/T$ ) along with sidebands adjacent to the harmonics due to the modulation. In the phasor representation of (166), every time-harmonic term is taken as the real part of the corresponding complex function which allows us to separate out the modula-

tion sidebands, e. g.

$$\cos(2\pi nt/T) = \text{Re} \{ e^{in\omega_0 t} \} \quad \text{CARRIER} \quad (167)$$

$$\cos 2\pi \left( \frac{n}{T} - f_{am} \right) = \text{Re} \{ e^{i(n\omega_0 + \omega_{am})t} \} \quad \text{UPPER AM SIDEBAND} \quad (168)$$

$$\cos 2\pi \left( \frac{n}{T} + f_{am} \right) = \text{Re} \{ e^{i(n\omega_0 - \omega_{am})t} \} \quad \text{LOWER AM SIDEBAND} \quad (169)$$

The *phasor amplitudes* are the complex coefficients on the harmonic term  $\exp(i\omega_0 t)$ . That is, the upper sideband has a phasor amplitude  $e^{+i\omega_{am} t}$  and the lower sideband has amplitude  $e^{-i\omega_{am} t}$ . The phasor picture provides a very useful way of viewing the modulation processes. If we plot the phasor amplitudes in the complex plane in the rotating reference frame of the carrier, we see immediately how the addition of the upper and lower AM sidebands to the carrier effect amplitude modulation. This is shown in Fig. 22a.

### Fourier Analysis of Sinusoidal Phase Modulation

We can now also analyze the phase modulation effect in a similar way. Let

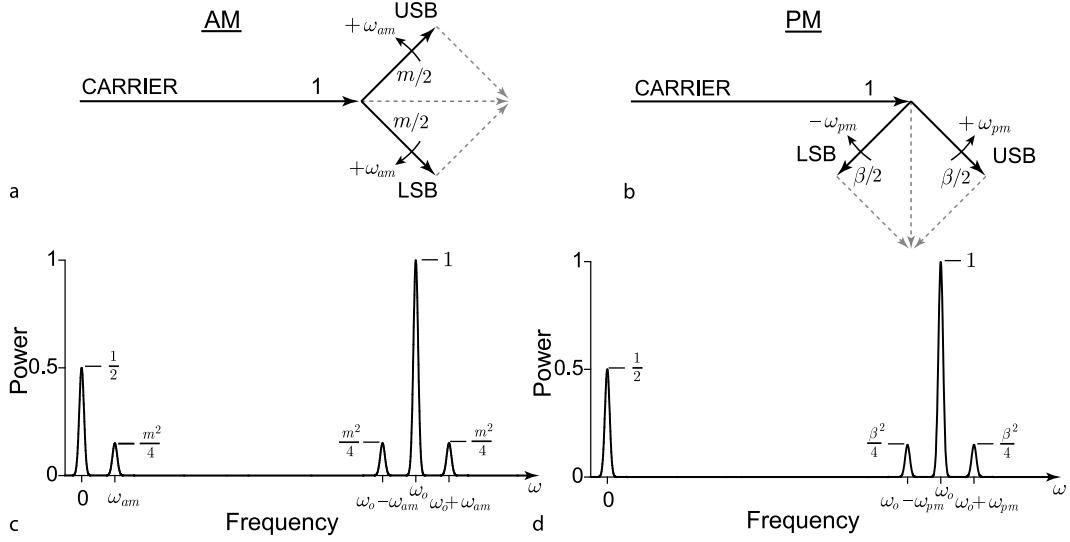
$$\begin{aligned} I(t) &= \sum_n I_o \\ &\cdot \exp \left[ -4 \ln 2 \left( \frac{t - [n + (\beta/2\pi) \cos(\omega_{pm} t)]T}{\tau} \right)^2 \right]. \quad (170) \end{aligned}$$

We wish to write a Fourier series for this representation but the actual period of the pulses is based on the modulation frequency,  $\omega_{pm}$  and not the repetition rate  $1/T$ . This makes the result cumbersome and difficult to work with. In almost all cases of interest, however, the modulation frequencies of the noise processes contributing to laser timing fluctuations are very slow compared with the pulse repetition rate, so we can approximate the time shift  $(\beta/2\pi) \cos(\omega_{pm} t)T$  as being static for the purposes of calculating the Fourier series. If we let  $\omega_{pm} t \equiv \Phi$  we may express  $I(t)$  in a complex Fourier series

$$I(t) = \sum_{n=-\infty}^{\infty} c_n e^{i2\pi(n/T)t} \quad (171)$$

where

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} I(t) e^{-i2\pi(n/T)t} dt \quad (172)$$



**Noise and Stability in Modelocked Soliton Lasers, Figure 22**

Relationships between sidebands and power spectrum for amplitude (AM) and phase (PM) modulated waveforms. **a** In amplitude modulation the upper (USB) and lower (LSB) modulation sidebands spinning at a rate  $\pm\omega_{am}$  align in phase with the carrier and their superposition adds and subtracts from the carrier amplitude. **b** In phase modulation the upper (USB) and lower (LSB) modulation sidebands spinning at a rate  $\pm\omega_{pm}$  align in quadrature with the carrier and their superposition periodically advances and retards the phase of the carrier. **c** Power spectrum of a pulse sequence undergoing AM only showing the baseband and fundamental spectral components. **d** Power spectrum of a pulse sequence undergoing PM only showing the baseband and fundamental spectral components. Note that there are no modulation spectral components at baseband and thus measurements of PM must be made about the fundamental  $\omega_o$  (or any harmonic)

$$= \frac{I_o}{T} \int_{-T/2}^{T/2} \exp \left[ -4 \ln 2 \left( \frac{t - (\beta T/2\pi) \cos \Phi}{\tau} \right)^2 \right] \cdot e^{-i2\pi(n/T)t} dt \quad (173)$$

$$= I_o \frac{\tau}{T} \sqrt{\frac{\pi}{4 \ln 2}} \exp \left( -\frac{(\pi(n/T)\tau)^2}{4 \ln 2} \right) e^{-in\beta \cos \Phi} \quad (174)$$

$$\equiv I'_o e^{-in\beta \cos \Phi} \quad (175)$$

Notice that the Gaussian term in (174) does not differ appreciably from unity until  $n \approx T/\tau$ . Since we seldom investigate harmonics of this high order, we can ignore the dependence on  $n$  and treat it as a constant. Substituting this result back into (171) and returning the value of  $\Phi = \omega_{pm}t$  yields

$$\begin{aligned} I(t) &= I'_o \sum_{n=-\infty}^{\infty} e^{i(2\pi(n/T)t - n\beta \cos \Phi)} \\ &= I'_o \sum_{n=-\infty}^{\infty} e^{i2\pi(n/T)t} e^{-in\beta \cos \omega_{pm}t} \end{aligned} \quad (176)$$

The second complex exponential can be rewritten in terms of the Bessel functions  $J_n(x)$  [77] as

$$e^{-in\beta \cos \omega_{pm}t} = \sum_{k=-\infty}^{\infty} (-i)^k J_k(n\beta) e^{ik\omega_{pm}t} \quad (177)$$

The Bessel functions are tabulated in many books [78] and have the following symmetry properties

$$J_k(\beta) = J_{-k}(\beta), \quad J_k(-\beta) = J_k(\beta); \quad n \text{ even} \quad (178)$$

$$J_k(\beta) = -J_{-k}(\beta), \quad J_k(-\beta) = -J_k(\beta); \quad n \text{ odd} \quad (179)$$

Thus

$$I(t) = I'_o \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} (-i)^k J_k(n\beta) e^{i(2\pi(n/T) + k\omega_{pm})t} \quad (180)$$

$$= I'_o \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} (-i)^k J_k(n\beta) e^{i(n\omega_o + k\omega_{pm})t} \quad (181)$$

where we have replaced  $2\pi/T$  with  $\omega_o$ , the radian fundamental frequency. This double series is actually easier to

interpret than it looks. The indices  $n$  and  $k$  set the oscillation frequencies of the time-harmonic function  $e^{i\omega(n,k)t}$ . Since  $1/T$  is the nominal repetition rate of the pulses, then  $n/T$  (or  $n\omega_0$ ) is the frequency of the  $n$ th harmonic of the main Fourier series. The slow phase modulation rate  $\omega_{pm}$  produces sidebands adjacent to the harmonics at frequencies  $n\omega_0 \pm k\omega_{pm}$ . Let's write out a few of the important terms in the series;

$$\begin{aligned}
 I(t) = & I'_0 \left\{ \dots \right. \\
 & + e^{-i2\omega_0 t} \left[ \dots - J_{-2}(2\beta)e^{-i2\omega_{pm}t} - iJ_{-1}(2\beta)e^{-i\omega_{pm}t} \right. \\
 & \quad \left. + J_0(2\beta) + iJ_1(2\beta)e^{i\omega_{pm}t} - J_2(2\beta)e^{i2\omega_{pm}t} + \dots \right] \\
 & + e^{-i\omega_0 t} \left[ \dots - J_{-2}(\beta)e^{-i2\omega_{pm}t} - iJ_{-1}(\beta)e^{-i\omega_{pm}t} \right. \\
 & \quad \left. + J_0(\beta) + iJ_1(\beta)e^{i\omega_{pm}t} - J_2(\beta)e^{i2\omega_{pm}t} + \dots \right] + 1 \\
 & + e^{i\omega_0 t} \left[ \dots - J_{-2}(\beta)e^{-i2\omega_{pm}t} + iJ_{-1}(\beta)e^{-i\omega_{pm}t} \right. \\
 & \quad \left. + J_0(\beta) - iJ_1(\beta)e^{i\omega_{pm}t} - J_2(\beta)e^{i2\omega_{pm}t} - \dots \right] \\
 & + e^{i2\omega_0 t} \left[ \dots - J_{-2}(2\beta)e^{-i2\omega_{pm}t} + iJ_{-1}(2\beta)e^{-i\omega_{pm}t} \right. \\
 & \quad \left. + J_0(2\beta) - iJ_1(2\beta)e^{i\omega_{pm}t} - J_2(2\beta)e^{i2\omega_{pm}t} - \dots \right] \\
 & \left. + \dots \right\}. \quad (182)
 \end{aligned}$$

We see that all of the phase modulation information is reproduced about each harmonic, but from a practical standpoint, the terms of greatest interest in the series are those associated with the DC or baseband ( $n = 0$ ) and the fundamental ( $n = \pm 1$ ). Notice that at baseband there is no modulation information. That is, the series returns a constant (+1) for  $n = 0$ . There is no spectral energy corresponding to the timing jitter near 0 Hz, but, the power in that DC term is proportional to the average laser power and thus is still useful for normalization purposes. In contrast, the fundamental and all harmonics have modulation sidebands and spectrum analyzers tuned to these components can determine the degree of timing jitter. Let's look a little more closely at the fundamental and its associated sidebands. Using the symmetry properties (178), (179) and

the small-argument approximations

$$J_0(x) \approx 1, \quad J_1(x) \approx x/2, \quad J_n(x) \approx 0, \quad |n| > 1 \quad (183)$$

we have

$$I_1(t) = I'_0 e^{i\omega_0 t} \left[ iJ_{-1}(\beta)e^{-i\omega_{pm}t} + J_0(\beta) - iJ_1(\beta)e^{i\omega_{pm}t} \right] \quad (184)$$

$$= I'_0 e^{i\omega_0 t} \left[ 1 - i\frac{\beta}{2} (e^{i\omega_{pm}t} + e^{-i\omega_{pm}t}) \right]. \quad (185)$$

Notice that, unlike the case of the AM sidebands in (168), (169), the PM sidebands are in phase-quadrature with the carrier, owing to the presence of the imaginary term  $i$ —preceding the sidebands. We can absorb this into the sideband phase and write

$$\begin{aligned}
 I_1(t) = & I'_0 \left[ \underbrace{e^{i\omega_0 t}}_{\text{CARRIER}} + \underbrace{\frac{\beta}{2} e^{i(\omega_{pm}t - \pi/2)}}_{\text{UPPER PM SIDE BAND}} + \underbrace{\frac{\beta}{2} e^{-i(\omega_{pm}t + \pi/2)}}_{\text{LOWER AM SIDE BAND}} \right]. \quad (186)
 \end{aligned}$$

A pictorial representation of the carrier and sideband phasors is shown in Fig. 22b. Notice that in the limit of very small modulation index ( $\beta \ll 2\pi$ ) the superposition of the upper and lower-sideband phasors only affect the phase angle of the carrier and not the amplitude since they are in phase-quadrature with respect to the carrier.

### Modulation Sideband Power

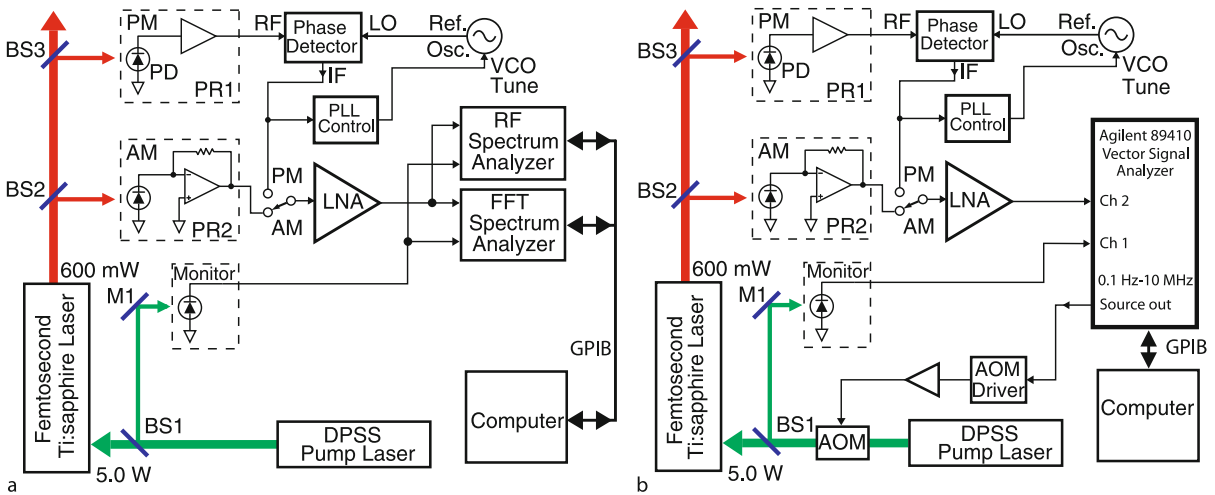
From a measurement standpoint, the most important attribute of the series expansions for AM and PM are the powers associated with the modulation sidebands. Because noise is a random process, we will be able to profitably discuss the power spectral density adjacent to each harmonic component in the series and we would like to be able to relate this to the relative magnitude of the noise modulation. This is most easily done by quantifying the ratio of the power in a single sideband to that in the carrier (usually the fundamental). Figure 22c shows the theoretical power contained in a purely AM-modulated periodic waveform of the type in Fig. 21a normalized to the fundamental carrier power. Figure 22d shows the same spectrum for the case of a purely PM-modulated periodic waveform of the type in Fig. 21b again normalized to the fundamental carrier power.

We can summarize the important relationships between the modulation types and their effects as in Table 5.

### Noise and Stability in Modelocked Soliton Lasers, Table 5

Principal results of AM and PM modulation applied to a periodic sequence of short laser pulses. The Fourier analysis of the pulse train is applied to the laser pulse *intensity* rather than the electric field since a photodetector produces a current that is linearly proportional to the laser power. The indicated power ratios are therefore those that would be observed on an RF spectrum analyzer connected to a photodiode which is illuminated by the optical pulse train

MODULATION TYPE	MODULATION INDEX	FREQUENCY	PEAK FLUCTUATION	POWER RATIOS	
				BASEBAND	CARRIER
AMPLITUDE (AM)	$m$	$\omega_{\text{am}}$	$\Delta I_{\text{PK}} = m I_0$	$\frac{P_{\text{mod}}}{P_{\text{avg}}} = \frac{m^2}{2}$	$\frac{P_{\text{SSB-AM}}}{P_c} = \frac{m^2}{4}$
PHASE (PM)	$\beta$	$\omega_{\text{pm}}$	$\Delta t_{\text{PK}} = \frac{\beta}{2\pi} T$	N/A	$\frac{P_{\text{SSB-PM}}}{P_c} = \frac{\beta^2}{4}$



### Noise and Stability in Modelocked Soliton Lasers, Figure 23

**a** System for measuring the amplitude and envelope phase noise of a modelocked Ti:sapphire laser and the amplitude noise of the pump laser. **b** Measurement system for characterizing the complex noise transfer function using a vector network analyzer in place of the spectrum analyzers. Legend: PR1; photoreceiver for phase noise, PR2; photoreceiver for amplitude noise, RF; radiofrequency input, LO; local oscillator or reference input, IF' intermediate (difference) frequency output, LNA; low noise amplifier, BS1, BS2, BS3; beamsplitters, M1; mirror, DPSS; diode-pumped solid-state pump laser, GPIB; general purpose interface bus, PLL; phase-locked loop, VCO; voltage-controlled oscillator, AOM; acousto-optic modulator

### Measurement of Laser Amplitude and Phase Noise

Making accurate, high dynamic range noise measurements of lasers can be challenging [79]. To accurately quantify the noise associated with amplitude and timing fluctuations of a periodic pulse train we rely on the use of Fast Fourier Transform (FFT) and radiofrequency (RF) spectrum analyzers applied to the detected photocurrent produced by a photodiode illuminated with the laser under study. These instruments are, essentially, tunable narrowband frequency filters that reveal the power spectral density contained in a narrow filter bandwidth at a specific frequency. One might be tempted to argue that using a photodiode and oscilloscope would be adequate but a quick calculation indicates that in order to see a wide-enough spectrum of noise on the oscilloscope, we are lim-

ited by thermal noise to observing a noise floor of several millivolts and hence the dynamic range is of order 60 dB (for a 1 volt peak signal). this amounts to an optical power fluctuation of approximately 0.1% which is enormous for a "well-behaved" laser. In fact, as we shall see, we need a dynamic range approaching 160 dB, or better, to learn anything meaningful about laser noise.

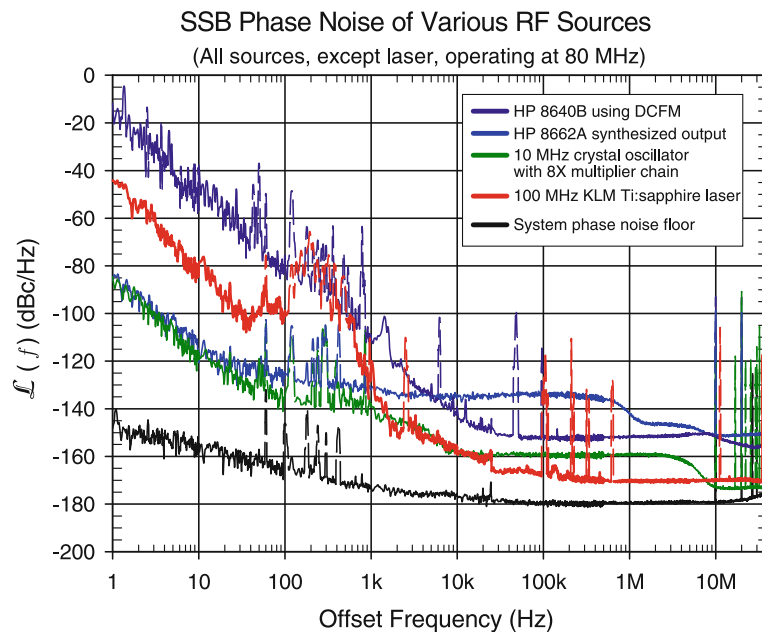
Figure 23a shows a schematic diagram of a system for measuring the amplitude and phase noise of the continuous train of pulses from a modelocked Ti:sapphire laser [79]. This laser is pumped from a diode-pumped solid-state laser (DPSS). In this system both the pump amplitude (AM) noise and the Ti:sapphire amplitude and phase (PM) noise are measured. The amplitude noise of either laser is manifested as fluctuations in the average power so only a relatively slow photodiode and receiver

are required. The amplitude noise is measured directly in the frequency domain by the spectrum analyzers. The FFT analyzer covers the frequency range from 0.1 Hz to 25 KHz and the RF spectrum analyzer from 10 Hz to 40 MHz. The frequency overlap allows accurate calibration and matching of noise spectra. To measure the phase noise of the pulse train envelope, we must compare its phase stability against a reference source. This is done using a high-quality reference oscillator in a phase-locked loop circuit. The photodiode and amplifier associated with this portion of the circuit (PR1) are tuned to the fundamental repetition rate of the laser (e.g. 100 MHz). The resulting sinewave is mixed against the reference oscillator sinewave and the output of the mixer ( $\approx 0$  Hz) is fed back to the tuning control of the reference oscillator. This forces the oscillator to stay phase-locked to the laser's repetition rate. For timing fluctuations that occur at a rate faster than the loop can respond, the voltage at the output of the mixer is linearly proportional to the phase offset and the power spectrum of these phase fluctuations are measured by the spectrum analyzers. In short, the phase detector circuit converts phase noise into amplitude noise which is characterized by the spectrum analyzers.

Figure 24 shows a comparison of phase noise between several conventional RF sources and a free-running modelocked Ti:sapphire laser. The displayed phase noise is displayed to represent the noise power spectral density of

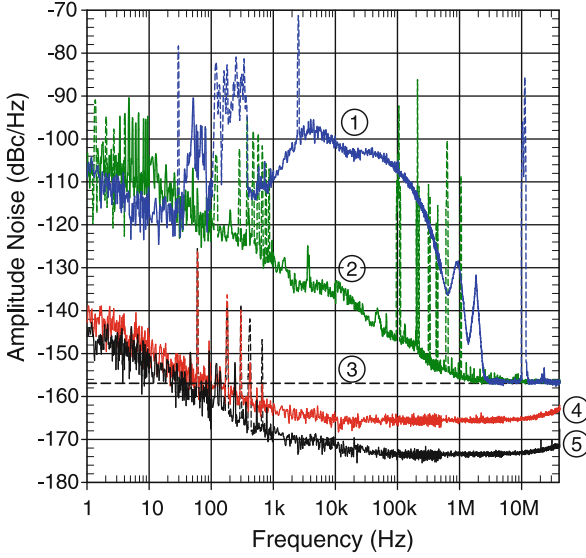
phase fluctuations on one side of the carrier referenced to the carrier power in decibels per-unit-bandwidth or dBc/Hz. The HP 8640B is a cavity-tuned signal generator whose frequency is steered via a varactor diode (so-called "DCFM"). The HP 8662A is an RF frequency synthesizer employing many complicated analog phase-locked loops to provide 1 Hz resolution through the range 10 KHz to 1300 MHz. The 10 MHz crystal oscillator is a high-performance, temperature compensated unit that is frequency-multiplied to 80 MHz. The Ti:sapphire laser used in this measurement was running at 100 MHz and thus will display a slightly higher phase noise for the same timing fluctuation as it would if it were running at 80 MHz, the difference scaling as  $(100/80)^2 = 1.56 = 1.94$  dB. From the data shown in a phase noise plot such as this, we can discern the average phase fluctuations at a given offset frequency of any source and compare them side-by-side. Modulation of the source phase, though a noise process, will show up as phase noise sidebands with magnitude given by the modulation index derived earlier in Subsect. "Modulation Sideband Power".

The format of the display of noise information is important and we should make a few remarks regarding this. Noise is a fundamentally statistical process and the noise that accompanies the laser electric field interferes with our ability to characterize it. By this we mean that given knowledge of the instantaneous electric field at any given mo-



Noise and Stability in Modelocked Soliton Lasers, Figure 24  
Comparison of phase noise performance of various sources





**Noise and Stability in Modelocked Soliton Lasers, Figure 25**  
Power spectral density of amplitude noise for two different pump lasers. 1 Argon-ion laser, 2 Diode-pumped neodymium-doped yttrium orthovanadate laser (Nd:YVO<sub>4</sub>). Both lasers producing 2.5 mA average photocurrent, 3 Calculated shot noise floor for  $i_{\text{avg}} = 2.5$  mA, 4 Photoreceiver noise floor ( $i_{\text{avg}} = 0$  mA), 5 Spectrum analyzer system noise floor. (From [79])

ment, we cannot say with certainty what the field will be at some specified time later. However, provided that the statistical properties of the noise are time-invariant, it can be shown that the *power spectrum* of the electric field will be stationary (unchanging). That is yet another reason why the spectrum analyzer is the appropriate tool for studying noise. It is, in effect, a tunable narrowband power meter, and thus the relevant characteristic of the noise is the power spectral density with units of Watts/Hertz. Relating the noise power-spectral-density to the fluctuating laser pulse train can best be understood using a deterministic model, which is valid on the time scale of the inverse bandwidth of the device measuring the fluctuation (in this case, the spectrum analyzer).

Figure 25 shows measurements of typical pump laser amplitude noise. The top curve (#1) is the noise spectrum of an argon-ion laser. The second curve (#2) is the noise spectrum of a diode-pumped, solid-state Nd:YVO<sub>4</sub> laser which is currently the preferred choice as a replacement pump source for pump Ti:sapphire lasers. Apart from the dramatic improvement in efficiency, this laser evidently has vastly superior noise performance. The photoreceiver and measurement system noise floors are shown in curves (#4) and (#5) respectively while the shot noise limit for the indicated photocurrent is shown as the straight line (#3).

### Pump-Induced Noise and the Noise Transfer Function

The amplitude stability of the pump laser plays a very important role in the amplitude and timing stability of the modelocked laser which it pumps [80,81,82,83,84,85]. Figure 26 shows the effect pictorially. Here a Kerr-lens modelocked Ti:sapphire laser is pumped by a diode-pumped solid-state laser (DPSS) with a highly exaggerated amount of amplitude noise. The Ti:sapphire laser emits a stream of femtosecond-duration pulses which have both amplitude noise and timing instability (also shown highly exaggerated).

We can understand how there might be a direct relationship between pump noise and the pumped-laser noise by considering several of the possible mechanisms that couple pump amplitude to the amplitude and repetition rate of a modelocked laser. For example, the pump power level (and hence pumping rate) establishes the gain of the laser and once above threshold, the power output is directly tied to the pumping rate  $r$  (70). Thus, amplitude noise on the pump beam will induce amplitude noise on the modelocked laser beam. We call this process AM-to-AM noise transfer [85]. The mechanism for this type of noise transfer can be easily understood for slow fluctuations. Figure 27 shows a simplified plot of the laser output power  $P_L$  versus pump power  $P_P$ . Above threshold the output power is linearly related to the pump power as derived in Sect. “Laser Dynamics” (Eq. (70)). For slow or static changes in pump power, the operating point just shifts along the straight line above threshold. We typically define a “slope efficiency”,  $\eta$ , of a laser above threshold as the slope of  $P_L$  vs.  $P_P$  at the operating point  $P_{P_0}$ . For the straight line shown in Fig. 27, we have

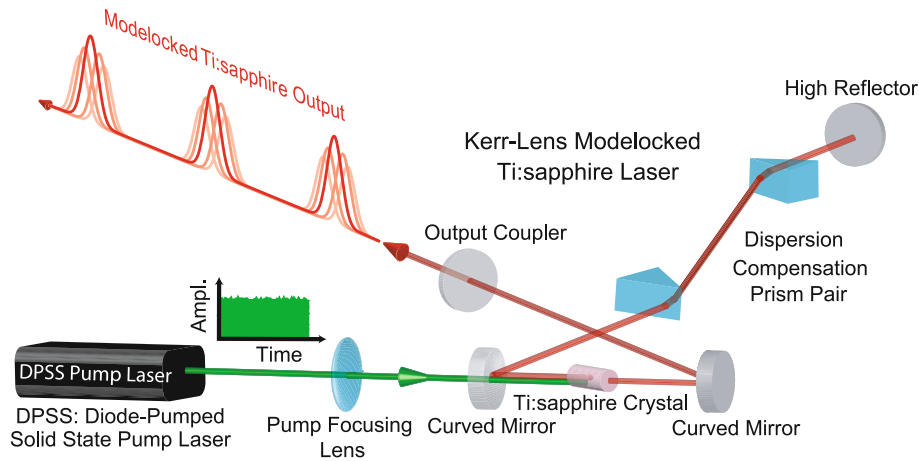
$$\eta = \left. \frac{dP_L}{dP_P} \right|_{P_P=P_{P_0}} = \frac{\Delta P_L}{\Delta P_P}. \quad (187)$$

If the pump modulation varies sinusoidally at a rate  $\omega_m$  low enough such that the laser dynamics follow perfectly, we can define a modulation index for both the pump and the laser in terms of the peak fractional deviations in the powers,

$$m_P \equiv \frac{\Delta P_P}{P_{P_0}}, \quad m_L \equiv \frac{\Delta P_L}{P_{L_0}}. \quad (188)$$

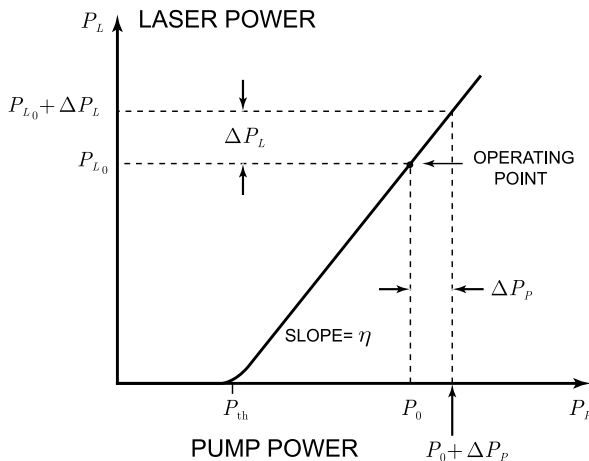
Let us further define the AM noise transfer function,  $H_{AM}(\omega_m)$  as the ratio of the induced modulation to the causative modulation. In the steady state ( $\omega_m = 0$ ), we see the direct relationship to the slope efficiency by

$$H_{AM}(0) \equiv \frac{m_L}{m_P} = \eta \frac{P_{P_0}}{P_{L_0}}. \quad (189)$$



**Noise and Stability in Modelocked Soliton Lasers, Figure 26**

Pump laser noise contributing to amplitude noise and timing instability in a modelocked laser



**Noise and Stability in Modelocked Soliton Lasers, Figure 27**

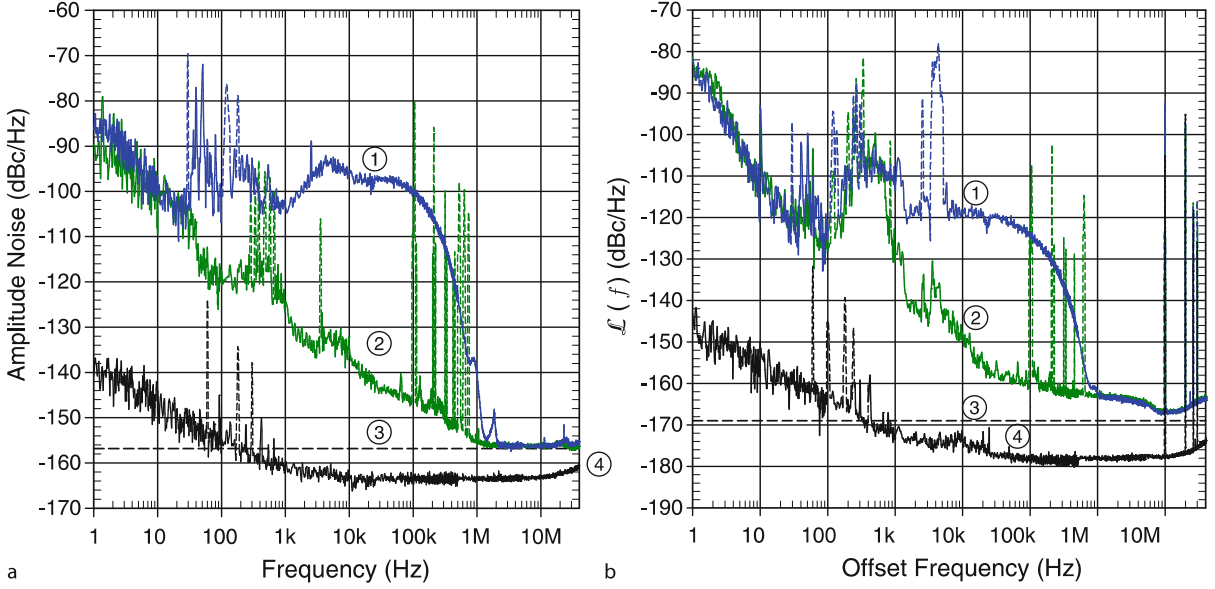
Laser output power  $P_L$  as a function of pump power  $P_P$  above threshold  $P_{th}$

This relationship will serve as the model for the AM-to-PM noise transfer function as well. That is, both noise transfer functions relate the induced effect (amplitude or timing fluctuation) to the pump modulation. In addition, (189) will also serve as a calibration or verification point for measurements of  $H_{AM}(\omega)$ , over a wide range of modulation frequencies, when the pump is intentionally modulated to simulate a noise component.

In addition to inducing amplitude noise in the modelocked laser, pump noise will also cause timing jitter or envelope phase noise in the modelocked pulse train. The relationship between the two is characterized by the phase noise transfer function  $H_{PM}(\omega)$ , which will be developed shortly. The sensitivity of the repetition-rate of the mod-

elocked laser to pump fluctuations has many potential contributing factors. Below threshold the population levels change in direct proportion to the pumping rate (67) but above threshold, when the laser is actually oscillating, the population inversion is clamped (69) and therefore the index of refraction is fixed. However, this is only true in the steady-state. As we have seen, if the pump is changed abruptly, the population levels cannot follow and take a while to settle down (Fig. 9). This means that for pump fluctuation frequencies that are high enough, but not substantially beyond the relaxation oscillation frequency, there can be modulation of the population levels and thus the index of refraction will also be modulated. This modulation of the index through the population inversion will cause timing jitter since the round-trip cavity time which establishes the pulse repetition rate depends on the optical path length and hence the index of refraction everywhere along the path.

Consider next the thermal loading of the laser crystal. Not every photon of absorbed pump power goes into inverting an atom. And, certainly, there is loss of energy in so-called non-radiative transitions as the atom cascades down from the highest pumped energy level to the ground state. In any case, there is a substantial amount of heat deposited into the laser crystal and often this necessitates water cooling. The equilibrium temperature achieved under steady-state pumping conditions will be many degrees above ambient room temperature and, therefore, fluctuations in the pumping power will cause fluctuations in the crystal temperature. The index of refraction of the laser crystal depends directly on the temperature through the thermo-optic coefficient which for Ti:sapphire has the value  $\partial n / \partial T = 1.2 \times 10^{-6} \text{ K}^{-1}$  [86]. Once again, a change



**Noise and Stability in Modelocked Soliton Lasers, Figure 28**

Evidence of noise transfer from pump laser. **a** Amplitude noise of Ti: sapphire laser pumped with 1 argon ion laser and 2 diode-pumped solid state laser. 3 Shot noise floor. 4 Photoreceiver noise floor. All spectra referenced to the average photocurrent. **b** Absolute single-sideband phase noise of the same Ti:sapphire laser with numbers corresponding to plots of AM spectra, left. All curves referenced to carrier amplitude. (From [79])

in the index of refraction will cause a change in the round-trip time delay for the pulses and manifest itself as timing jitter. Along with the direct thermo-optic effect, we have a thermally-driven expansion of the laser crystal at a differential rate of  $5.8 \times 10^{-6} \text{ K}^{-1}$  which also increases the cavity round trip time.

Yet another contribution to timing jitter comes about in modelocked lasers running high peak-power pulses. The index of refraction has a nonlinear contribution due to the high peak power  $I(t)$  such that  $n = n_0 + n_2 I(t)$  where  $n_2 = 3.45 \times 10^{-16} \text{ cm}^2/\text{W}$  [11]. In addition, when the index of refraction is modulated, the angle of refraction at the crystal surfaces due to Snell's law changes and the beam path through the dispersion compensation prism pair is modified slightly (see Fig. 26). The differential variation in time delay for a typical Kerr-lens modelocked Ti:sapphire laser due to an index change is  $d\tau/dn \approx 4 \times 10^{-11} \text{ s}$ .

Figure 28 shows directly the effect of pump noise inducing AM and PM noise on the pulse train of modelocked Ti:sapphire laser for two different types of pump laser. On the left is the amplitude noise of the Ti:sapphire laser measured at baseband and on the right is the single-sideband phase noise measured about the carrier at the fundamental repetition rate of 80 MHz. Trace #1 is the noise for the case of an argon-ion pump laser and trace #2

is for a diode-pumped solid-state (DPSS) single-frequency laser. Notice the dramatic improvement in noise reduction using the DPSS pump laser. Notice, also, the direct correlation between the noise spectra of the pump lasers, as shown in Fig. 25, and the noise spectra of the laser that is being pumped (Fig. 28). The relationship between the noise spectra of the pump lasers and the pumped laser can be characterized by the noise transfer functions  $H_{AM}(\omega)$  and  $H_{PM}(\omega)$ .

**Analytical Treatment of Amplitude-to-Amplitude Noise Transfer** In contrast to the large-scale fluctuations characteristic of turn-on transients and routes to chaos, pump-power noise is expected to be very small in amplitude and thus a perturbation approach applied to the coupled cavity-atom rate equations is quite appropriate. We assume steady-state operation above threshold and therefore rewrite those equations without the spontaneous emission term

$$\frac{dn(t)}{dt} = KN(t)n(t) - \gamma_c n(t) \quad (190)$$

$$\frac{dN(t)}{dt} = R_p(t) - \gamma_2 N(t) - KN(t)n(t). \quad (191)$$

As we saw previously, the population inversion is clamped at threshold,  $N_{th} = \gamma_c/K$  (69), and the steady-

state photon density is given by (70)

$$n = n_{ss} = \left( \frac{R_{P0}}{\gamma_2 N_{th}} - 1 \right) \frac{\gamma_2}{K} = (r - 1) \frac{\gamma_2}{K}. \quad (192)$$

The goal now is to find out how the population inversion and photon density respond to a small sinusoidal variation in the pumping rate. Before we apply this small-signal perturbation to the pump, we first find the natural response of the system (190), (191) to perturbations about the steady-state. We assume that  $n(t)$  and  $N(t)$  take the forms

$$\begin{aligned} n(t) &= n_{ss} + n_1(t), & n_1(t) &\ll n_{ss} \\ N(t) &= N_{th} + N_1(t), & N_1(t) &\ll N_{th} \end{aligned} \quad (193)$$

and substitute them into (190), (191) to obtain the following linearized small-signal rate equations

$$\frac{dn_1(t)}{dt} = (r - 1)\gamma_2 N_1(t) \quad (194)$$

$$\frac{dN_1(t)}{dt} = -\gamma_c n_1(t) - r\gamma_2 N_1(t). \quad (195)$$

In the usual way, we assume solutions that vary as  $e^{st}$  and obtain an equation for  $s$  with roots

$$s = s_1, s_2 = -\frac{r\gamma_2}{2} \pm \sqrt{\left(\frac{r\gamma_2}{2}\right)^2 - (r - 1)\gamma_2\gamma_c}. \quad (196)$$

Now, in the case of solid-state lasers where  $\gamma_2 \ll \gamma_c$ ,  $(r - 1)\gamma_2\gamma_c > (r\gamma_2/2)^2$ , and the roots form a complex conjugate pair

$$s_1, s_2 = -\frac{r\gamma_2}{2} \pm i\sqrt{(r - 1)\gamma_2\gamma_c - \left(\frac{r\gamma_2}{2}\right)^2}. \quad (197)$$

When we combine both complex solutions for  $n_1(t)$  corresponding to  $s_1, s_2$  with arbitrary coefficients, we obtain the real, damped, oscillatory behavior for the cavity photon number

$$n_1(t) = n_{o1}e^{s_1 t} + n_{o2}e^{s_2 t} = n_1 e^{-\gamma_{sp} t} \cos \omega'_{sp} t \quad (198)$$

where

$$\begin{aligned} \gamma_{sp} &\equiv \frac{r\gamma_2}{2}, & \omega'_{sp} &\equiv \sqrt{\omega_{sp}^2 - \gamma_{sp}^2}, \\ \omega_{sp} &\equiv \sqrt{(r - 1)\gamma_2\gamma_c}. \end{aligned} \quad (199)$$

The phase shift corresponding to initial conditions has been arbitrarily set equal to zero. The term  $\omega_{sp}$  is often referred to as the “spiking” or “relaxation oscillation” frequency as is readily apparent in the plots of the turn-on

transient solutions shown in Fig. 9. Thus, the solution to the first of (193) has the form

$$n(t) = n_{ss} + n_1 e^{-\gamma_{sp} t} \cos \omega'_{sp} t, \quad (200)$$

Next, to solve for  $N_1(t)$  we substitute (198) for  $n_1(t)$  back into (194) and (195) and find

$$N_1(t) = -2n_1 \frac{\gamma_c}{\omega_{sp}} e^{-\gamma_{sp} t} \sin(\omega'_{sp} t + \phi) \quad (201)$$

where

$$\phi = \tan^{-1} \left[ \frac{\gamma_{sp}}{\omega'_{sp}} \right]. \quad (202)$$

Finally, the complete solution for the population inversion is

$$N(t) = N_{th} - 2n_1 \frac{\gamma_c}{\omega_{sp}} e^{-\gamma_{sp} t} \sin(\omega'_{sp} t + \phi). \quad (203)$$

Note that since  $\omega'_{sp} \gg \gamma_{sp}$ , the population inversion *leads* the cavity photon number, as expected intuitively.

Now we can apply a small sinusoidal perturbation (modulation) to the pumping rate and solve for the effects on the photon and population inversion densities. Let

$$R_P(t) = R_{P0} + \text{Re}\{R_{P1} e^{i\omega_m t}\} s^{-1} m^{-3} \quad (204)$$

where  $R_{P1} \ll R_{P0}$ . Thinking ahead to a link with the noise transfer function, we can also write this in terms of a small-signal pump modulation index,  $m_p$ ,

$$R_P(t) = R_{P0} \left[ 1 + \text{Re}\{m_p e^{i\omega_m t}\} \right] \quad (205)$$

where

$$m_p \equiv \frac{R_{P1}}{R_{P0}}. \quad (206)$$

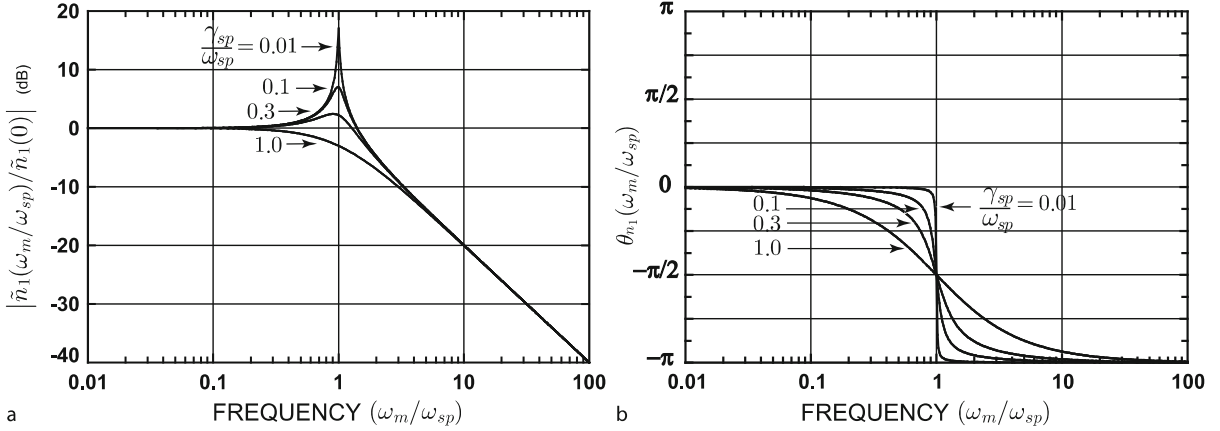
When the pump is modulated sinusoidally, we expect that our linearized system will yield sinusoidally varying responses. Therefore, we seek solutions of the form

$$n(t) = n_{ss} + \text{Re}\{\tilde{n}_1 e^{i\omega_m t}\}, \quad |\tilde{n}_1| \ll n_{ss} \quad (207)$$

$$N(t) = N_{th} + \text{Re}\{\tilde{N}_1 e^{i\omega_m t}\}, \quad |\tilde{N}_1| \ll N_{th} \quad (208)$$

where  $\tilde{n}_1$  and  $\tilde{N}_1$  are complex phasor amplitudes. When these expressions are substituted back into (190) and (191) we obtain complex phasor amplitudes for the cavity photon number density

$$\tilde{n}_1 = \frac{\omega_{sp}^2 R_{P1} / \gamma_c}{\omega_{sp}^2 - \omega_m^2 + i2\gamma_{sp}\omega_m} \quad (209)$$



**Noise and Stability in Modelocked Soliton Lasers, Figure 29**

Normalized magnitude (a) and phase (b) of the cavity photon density response for several values of  $\gamma_{sp}/\omega_{sp}$

and the population inversion density, respectively

$$\tilde{N}_1 = \frac{i\omega_m R_{P_1}}{\omega_{sp}^2 - \omega_m^2 + i2\gamma_{sp}\omega_m}. \quad (210)$$

It is very useful to consider plots of these expressions as a function of the normalized frequency  $\omega = \omega_m/\omega_{sp}$ . It is also helpful to separate them into magnitude and phase. After a little algebra we find

$$|\tilde{n}_1| = \frac{R_{P_1}}{\gamma_c} \frac{1}{\sqrt{\left(1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2\right)^2 + \left(2\frac{\gamma_{sp}}{\omega_{sp}} \frac{\omega_m}{\omega_{sp}}\right)^2}} \quad (211)$$

$$\theta_{n_1} = \tan^{-1} \left[ -\frac{2\frac{\gamma_{sp}}{\omega_{sp}} \frac{\omega_m}{\omega_{sp}}}{1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2} \right] = \tan^{-1} \left[ -\frac{2\gamma_{sp}\omega_m}{\omega_{sp}^2 - \omega_m^2} \right] \quad (212)$$

$$|\tilde{N}_1| = \frac{R_{P_1}}{\omega_{sp}} \frac{\omega_m/\omega_{sp}}{\sqrt{\left(1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2\right)^2 + \left(2\frac{\gamma_{sp}}{\omega_{sp}} \frac{\omega_m}{\omega_{sp}}\right)^2}} \quad (213)$$

$$\theta_{N_1} = \tan^{-1} \left[ \frac{1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2}{2\frac{\gamma_{sp}}{\omega_{sp}} \frac{\omega_m}{\omega_{sp}}} \right] = \tan^{-1} \left[ \frac{\omega_{sp}^2 - \omega_m^2}{2\gamma_{sp}\omega_m} \right]. \quad (214)$$

These functions, in normalized form, are shown in Figs. 29 and 30 for various values of the normalized damping term  $\gamma_{sp}/\omega_{sp}$ . They form the basis for understanding much of the behavior of the noise transfer functions that will be derived in following sections.

Now that we have expressions for the phasor amplitudes of the photon and population inversion densities, we

can use them in the definitions of the noise transfer functions. We define the complex AM-to-AM noise transfer function at modulation frequency  $\omega_m$  as the ratio of the induced laser amplitude modulation index to the pump modulation index

$$H_{AM}(\omega_m) \equiv \frac{\tilde{m}_L(\omega_m)}{m_p} = \frac{\tilde{n}_1(\omega_m)/n_{ss}}{R_{P_1}/R_{P_0}} \quad (215)$$

$$= \left( \frac{R_{P_0}}{n_{ss}R_{P_1}} \right) |\tilde{n}_1| e^{i\theta_{n_1}} \quad (216)$$

$$= |H_{AM}(\omega_m)| e^{i\theta_{AM}} \quad (217)$$

where

$$|H_{AM}(\omega_m)| \equiv \frac{R_{P_0}/n_{ss}\gamma_c}{\sqrt{\left(1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2\right)^2 + \left(2\frac{\gamma_{sp}}{\omega_{sp}} \frac{\omega_m}{\omega_{sp}}\right)^2}} \quad (218)$$

and

$$\theta_{AM} = \theta_{n_1} = \tan^{-1} \left[ -2\frac{\gamma_{sp}\omega_m}{\omega_{sp}^2 - \omega_m^2} \right]. \quad (219)$$

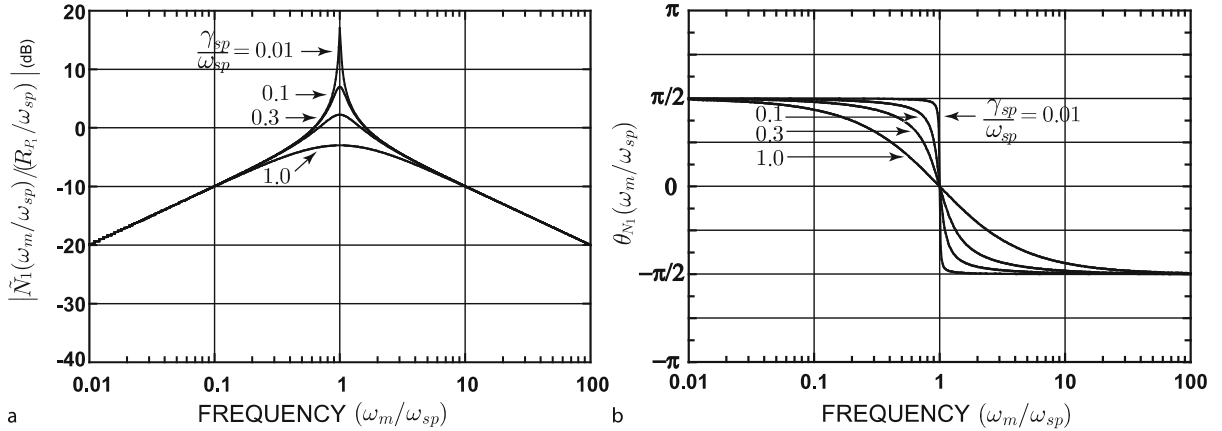
Apart from the DC response, we see that  $H_{AM}(\omega_m)$  has exactly the same form as the photon density plotted in Fig. 29.

The low frequency asymptotic form, or  $H_{AM}(0)$ , can be cast in a particularly revealing form by using (69) and (70);

$$H_{AM}(0) = \frac{R_{P_0}}{n_{ss}\gamma_c} = \frac{R_{P_0}K}{\gamma_c\gamma_2(r-1)} = \frac{R_{P_0}/N_{th}\gamma_2}{r-1} = \frac{r}{r-1}. \quad (220)$$

This expression tells us that, since the pumping rate is, by definition,  $r > 1$  above threshold, the induced AM noise





**Noise and Stability in Modelocked Soliton Lasers, Figure 30**

Normalized magnitude (a) and phase (b) of the population inversion response for several values of  $\gamma_{sp}/\omega_{sp}$

in the pumped laser at low frequencies *always exceeds the noise of the pump!*

The other two important attributes of the spectral dependence of  $H_{AM}(\omega_m)$  are the peaking at the relaxation oscillation (or spiking) frequency,  $\omega_{sp}$ , and the slope for  $\omega_m > \omega_{sp}$ . In the classic fashion of the damped harmonic oscillator, the cavity photon number, and hence  $H_{AM}(\omega_m)$  have a peak in the response at the resonance frequency with the magnitude of the peaking dependent upon the degree of damping,  $\gamma_{sp}/\omega_{sp}$ . Above the resonance, the system cannot follow the rapid oscillations in the pumping rate and  $H_{AM}(\omega_m)$  decreases at a rate of 20 dB/decade which is characteristic of the “second-order pole” in (211).

**Analytical Treatment of Amplitude-to-Phase Noise Transfer** As alluded to previously, the sources and mechanisms of amplitude-to-phase noise transfer are diverse and complicated. A full treatment is beyond the scope of the present discussion but we can get a glimpse of how to include all of the physical effects by beginning with a heuristic model and then establishing links between the model elements and the physical mechanisms.

Suppose we follow a packet of photons around the laser cavity and designate the round trip time as  $\tau_c$ . Then, if  $\tau_c$  is constant, it is reasonable to describe the train of optical pulses exiting the laser by

$$P(t) = P_o \sum_n A(t - n\tau_c) \quad (221)$$

where  $P_o$  is the peak optical power and  $A$  is the shape of an individual pulse. Since the noise processes that disturb the round trip cavity, and thus lead to timing fluctuations, are random, we will study the effects in the frequency domain

using the phase noise measurement apparatus described earlier. We can analyze one particular spectral component of the power spectral density of the phase fluctuations by including in (221) a sinusoidal variation at frequency  $\omega_m$  so that

$$P(t) = P_o \sum_n A\left(t - \tau_{co} \left[ n + \frac{\beta(\omega_m)}{2\pi} \cos(\omega_m t + \phi(\omega_m)) \right] \right) \quad (222)$$

where  $\tau_{co}$  is the unperturbed round-trip cavity time,  $\beta(\omega_m)$  is the peak phase deviation, and  $\phi(\omega_m)$  is a frequency-dependent phase offset between the output pulse train and whatever mechanism is causing the fluctuation. When expressed in this fashion, a peak deviation of  $\beta = 2\pi$  radians corresponds to one period of the pulse train, or one round-trip cavity time. Comparing (221) and (222) we see that the instantaneous round trip cavity time with the sinusoidal fluctuation can be written

$$\tau_c(t) = \tau_{co} + \tau_{co} \frac{\beta(\omega_m)}{2\pi} \cos(\omega_m t + \phi(\omega_m)). \quad (223)$$

This is a phenomenological description since it just represents the *effect* of the perturbation and not the cause. The cause of the perturbations can be included in a *physical* description in terms of propagation lengths and pulse group velocities;

$$\tau_c = 2 \left[ \frac{L_{ca}}{v_{air}} + \frac{L_{cp}}{v_{pr}} + \frac{L_a}{v_{ga}} \right]. \quad (224)$$

The first term accounts for the propagation time through air between the optical components in the cavity.  $L_{ca}$  is the length of the path and  $v_{air}$  is the velocity. The second term

accounts for the group delay through the prisms, if intra-cavity prisms are used for dispersion compensation. Finally, the third term accounts for the group delay through the active atomic medium.  $L_a$  is the length of the gain medium and  $v_{ga}$  is the group velocity through the medium. Variations in the pump power will cause variations in all quantities except the velocity in air,  $v_{air}$ . The variations in the remaining five are caused by the pump modulating the photon density, the population inversion and the laser rod temperature. We have analyzed the link between pump modulation and the first two earlier. We can build on this by once again assuming a sinusoidal pump fluctuation

$$R_P(t) = R_{P_0} + R_{P_1} \cos \omega_m t. \quad (225)$$

Once again we seek solutions of the form

$$n(t) = n_{ss} + \text{Re} \left\{ \tilde{n}_1 e^{i\omega_m t} \right\} \quad (226)$$

$$N(t) = N_{th} + \text{Re} \left\{ \tilde{N}_1 e^{i\omega_m t} \right\} \quad (227)$$

$$T(t) = T_o + \text{Re} \left\{ \tilde{T}_1 e^{i\omega_m t} \right\}. \quad (228)$$

The phasor amplitudes for  $\tilde{n}_1$  and  $\tilde{N}_1$  have been found (209), (210) and the laser rod temperature can be shown to respond as [87]

$$\tilde{T}_1(r') = \frac{2}{k} \frac{R_{P_1}}{R_{P_0}} \sum_{l=1}^{\infty} \frac{J_0(\beta_l r'/b)}{\beta_l^2 J_1^2(\beta_l)} \frac{1}{1 + i\omega_m \frac{b^2}{k\beta_l^2}} \quad (229)$$

$$= \frac{2}{k} \frac{R_{P_1}}{R_{P_0}} \sum_{l=1}^{\infty} \frac{J_0(\beta_l r'/b)}{\beta_l^2 J_1^2(\beta_l)} \frac{1}{\sqrt{1 + \left(\omega_m \frac{b^2}{k\beta_l^2}\right)^2}} e^{i\theta_T}, \quad (230)$$

$$\theta_T \equiv \tan^{-1} \left[ -\omega_m \frac{b^2}{k\beta_l^2} \right] \quad (231)$$

where  $k$  is the diffusivity ( $m^2/s$ ),  $\beta_l$  are the eigenvalues of the radial heat diffusion problem,  $r'$  is the radial coordinate in the laser rod and  $b$  is the radius of the rod.

By carefully considering the contributions of each of the fluctuating quantities  $\tilde{n}_1$ ,  $\tilde{N}_1$  and  $\tilde{T}_1$  to the total round-trip cavity time, we can gather the physical coupling factors into coefficients on the phasor amplitudes by writing

$$\tau_c(t) = A + B \text{Re} \left\{ \tilde{n}_1 e^{i\omega_m t} \right\} + C \text{Re} \left\{ \tilde{N}_1 e^{i\omega_m t} \right\} + D \text{Re} \left\{ \tilde{T}_1 e^{i\omega_m t} \right\}. \quad (232)$$

The derivation of the coupling coefficients is rather involved and will be presented elsewhere [88] but we note

the results here

$$A \equiv \tau_{co}(n_{go}) + \frac{2L_{a0}}{c} \left[ n_{host}(\omega_a) + n_2 I_{pk} + \omega_a \frac{dn_{host}}{d\omega} + \frac{e^2}{n_{host} m \epsilon_0 \Delta \omega_a^2} N_{th} \right] \quad (233)$$

$$B \equiv \frac{n_2 I_{pk}}{n_{ss}} \left[ \frac{d\tau_p}{dn} + 2 \frac{L_{a0}}{c} \right] \quad (234)$$

$$C \equiv \frac{2L_{a0}}{c} \frac{e^2}{n_{host} m \epsilon_0 \Delta \omega_a^2} \quad (235)$$

$$D \equiv \frac{d\tau_p}{dn} \frac{\partial n}{\partial T} + \alpha_T \frac{2L_{a0}}{c} \left[ n_{host}(\omega_a) + n_2 I_{pk} + \omega_a \frac{dn_{host}}{d\omega} + \frac{e^2}{n_{host} m \epsilon_0 \Delta \omega_a^2} N_{th} + \frac{1}{\alpha_T} \frac{\partial n}{\partial T} \right]. \quad (236)$$

The terms in these expressions are defined in the beginning of this chapter (Nomenclature) with the following exceptions

$\frac{d\tau_p}{dn}$  = slope of the group delay through prisms,

$\frac{\partial n}{\partial T}$  = thermo-optic coefficient.

The coupling mechanisms between the fluctuating quantities  $\tilde{n}_1$ ,  $\tilde{N}_1$  and  $\tilde{T}_1$  and the physical characteristics of the laser were discussed in the beginning of this section and are summarized in Table 6 below.

We now define the phase noise transfer function as the ratio of the peak induced complex phase deviation to the pump modulation index

$$H_{PM}(\omega_m) \equiv \frac{\tilde{\beta}(\omega_m)}{m_p} \quad (237)$$

**Noise and Stability in Modelocked Soliton Lasers, Table 6**

**Coupling between the modulated laser parameters (photon density, population inversion density and laser rod temperature) and the physical characteristics that determine the round-trip cavity time  $\tau_c$**

	PHOTON DENSITY $\tilde{n}_1$	POPULATION INVERSION DENSITY $\tilde{N}_1$	LASER ROD TEMPERATURE $\tilde{T}_1$
GAIN MEDIUM (LASER ROD) LENGTH			$D$
GAIN MEDIUM GROUP VELOCITY	$B$	$C$	$D$
CAVITY LENGTH (BEAM STEERING)	$B$		$D$

where the complex phase deviation  $\tilde{\beta}(\omega_m) \equiv \beta(\omega_m) e^{i\phi(\omega_m)}$ . Comparing (223) with (232) and using  $m_p = R_{P1}/R_{P0}$ , we have

$$H_{PM}(\omega_m) = \frac{2\pi R_{P0}}{\tau_{co} R_{P1}} \left( B \tilde{n}_1(\omega_m) + C \tilde{N}_1(\omega_m) + D \tilde{T}_1(\omega_m) \right) \quad (238)$$

where we have explicitly emphasized the frequency dependence of the phasor amplitudes.

### Measurement of the Complex Noise Transfer Function

A system for measuring the complex noise transfer function (NTF) is conceptually straightforward and quite similar to that used for laser noise measurements. Figure 23a shows a schematic diagram of such a system. The principal differences between the noise measurement system (Fig. 23b) and the NTF characterization system is the introduction of an acousto-optic amplitude modulator (AOM) into the pump beam path to simulate noise fluctuations and a vector signal analyzer to provide a stimulus and measure the induced response. The vector signal analyzer provides a swept sinusoidal signal from 0.1 Hz to 10 MHz and records the depth of pump modulation on Channel 1 and the depth of Ti:sapphire laser modulation on Channel 2. By switching between two specialized receivers, either the AM or PM response can be measured both in magnitude and phase compared with the pump signal. It is the complex ratio of the induced AM or PM from the Ti:sapphire laser to that of the pump that defines the AM- or PM-NTF, respectively.

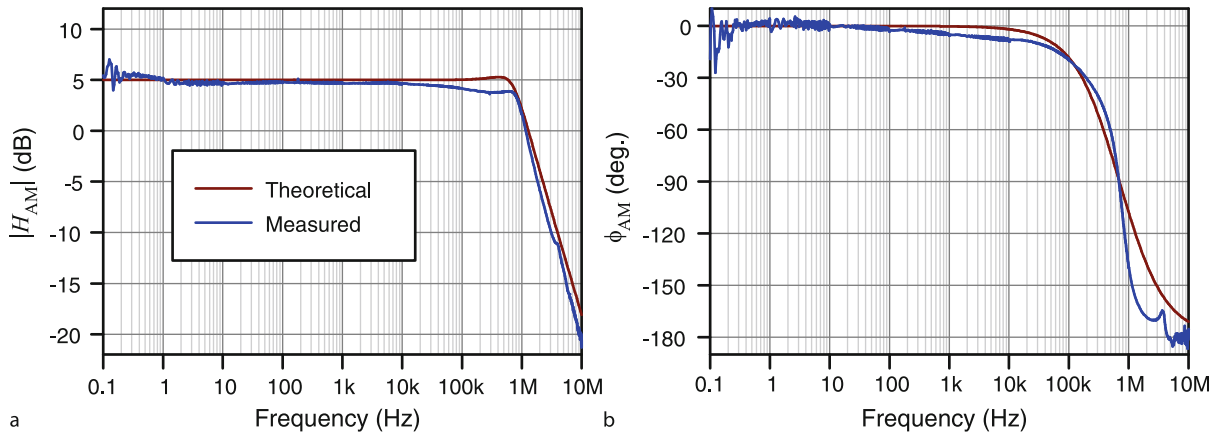
Figures 31 and 32 show the data for measurements of  $H_{AM}(\omega_m)$  and  $H_{PM}(\omega_m)$ , along with theoretical plots

based on (215) and (238) [85]. Since the AM noise transfer function,  $H_{AM}(\omega_m)$ , depends only on the photon density, it has exactly the same shape as the plot of  $\tilde{n}_1(\omega_m)$  shown in Fig. 29. There is a small amount of peaking at the relaxation oscillation frequency ( $\approx 500$  KHz) and sharp rolloff ( $-20$  dB/decade) thereafter. In the plots of  $H_{PM}(\omega_m)$  (Fig. 32) we see the composite behavior predicted by expression (238), that is, a superposition of the frequency-domain behavior of modulations of  $n$ ,  $N$  and  $T$ . At low frequencies, where thermal effects dominate, there is a  $-10$  dB/decade rolloff consistent with the first-order pole of (229). In the middle and upper frequency ranges ( $1 \text{ KHz} < f < 10 \text{ MHz}$ ) the behavior mimics that of the photon fluctuation  $\tilde{n}_1(\omega_m)$ . The population inversion density,  $\tilde{N}_1(\omega_m)$ , depicted in Fig. 30, with its characteristic  $\pm 10$  dB/decade slopes on either side of the relaxation oscillation frequency, does not appear because the magnitude of the effect of the coupling to the group velocity in the gain medium is swamped by the effects of the photon density.

### Predicting Amplitude and Phase Noise from the Pump Noise Spectrum and the Noise Transfer Function

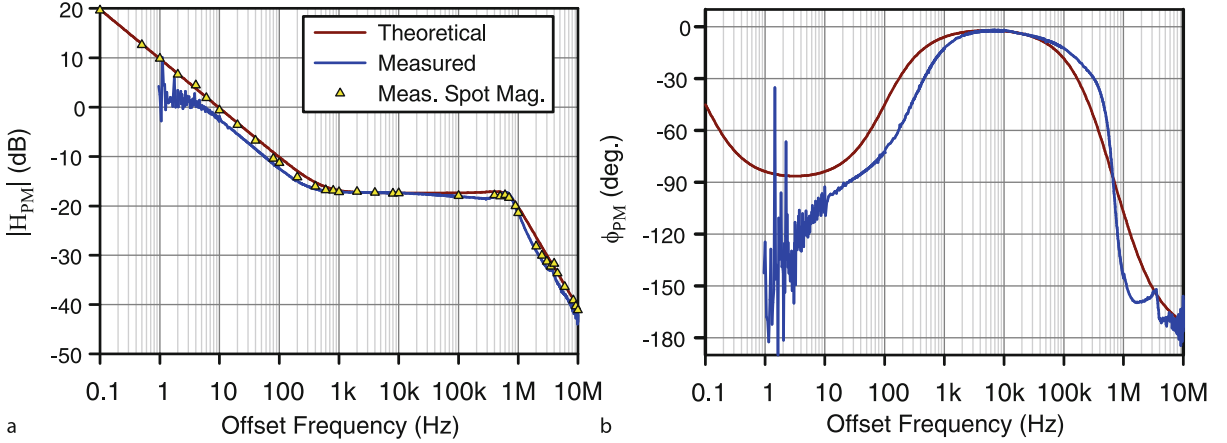
Since the noise transfer function represents the response of a linear system to the noise spectrum of a pump laser, we should be able to use it to predict the noise properties of the pumped laser once we have characterized the noise spectrum of the pump [89]. Thus, if  $S_P(\omega_m)$  is the noise power spectral density (PSD) of the pump laser (as measured in the photodetected current spectrum), the amplitude noise power spectral density is given by

$$S_{AM}(\omega_m) = |H_{AM}(\omega_m)|^2 S_P(\omega_m) \quad (239)$$



Noise and Stability in Modelocked Soliton Lasers, Figure 31

a Magnitude and b phase of the measured and theoretical complex AM NTF response. Continuous frequency measurements made with vector signal analyzer in setup shown in Fig. 23b



**Noise and Stability in Modelocked Soliton Lasers, Figure 32**

**a** Magnitude and **b** phase of the measured and theoretical complex PM NTF response. Continuous frequency measurements made with vector signal analyzer in setup shown in Fig. 23b. Yellow triangles are data taken at single frequency points (spot measurements) for verification of swept frequency approach and for better signal-to-noise below 5 Hz

and the single-sideband phase noise PSD is similarly given by

$$S_{\text{PM-SSB}}(\omega_m) = \frac{|H_{\text{PM}}(\omega_m)|^2}{2} S_p(\omega_m) \quad (240)$$

where the factor of two comes about from the phase noise being specified as the single-sideband portion about the carrier.

Figure 33a shows the noise spectra of two types of pump lasers that can be used for pumping modelocked Ti:sapphire lasers. The multi-wavelength argon ion laser (Coherent Innova 300C) clearly has much more noise than the single longitudinal-mode diode-pumped solid-state laser (DPSS). The DPSS laser in this case was a frequency-doubled Nd:YVO<sub>4</sub> laser (Coherent Verdi). Figure 33b shows the absolute value of noise transfer function of a 100 MHz Ti:sapphire laser measured using both of these pump sources. The very slight difference in the data is due to the different effective pumping powers and how that affects the relaxation oscillation. When the noise spectra of the pump lasers is multiplied by the modulus-squared NTF (239), (240), one obtains the predicted Ti:sapphire laser noise as is shown in Figs. 34 and 35. The correlation between the measured and the predicted spectra is very good and validates the concept of the noise transfer function method.

Finally, it is interesting to do a comparison between the Ti:sapphire noise spectra when pumped by both lasers and this is shown in Fig. 36. Notice the considerable improvement in both amplitude and phase noise when using the DPSS pump laser.

### Noise in Soliton Lasers

Much of the early work on instability in soliton lasers was developed by Haus and coworkers at M.I.T. In the following section we follow closely the pioneering approach presented by Haus and Mecozzi in their now-classic paper [65].

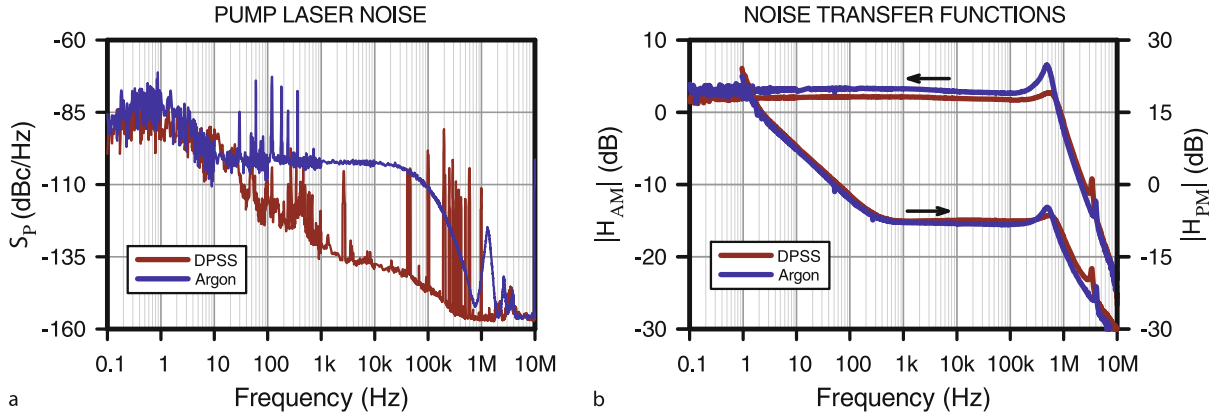
To study the issue of stability in soliton lasers, we return to the master equation (107) with a slight modification to allow for the fact that the pulse amplitude is not necessarily stable upon a cavity round trip [65]. Thus,

$$T_R \frac{\partial a(T, t)}{\partial T} = \left[ -l + g \left( 1 - \frac{1}{\Omega_g} \frac{\partial}{\partial t} + \frac{1}{\Omega_g^2} \frac{\partial^2}{\partial t^2} \right) + iD \frac{\partial^2}{\partial t^2} + (\gamma - i\delta)|a|^2 \right] a(T, t) + T_R S(t, T) \quad (241)$$

where  $a(T, t)$  is the field amplitude which is now both a function of the local short-time variable  $t$ , and the long-time variable  $T$ , on the order of many cavity round-trip times,  $T_R$ .  $a(T, t)$  is normalized such that  $|a|^2$  is equal to the instantaneous envelope power. The other parameters are the same as described in Sect. “Modelocked Soliton Lasers” except for the new term,  $S(t, T)$ , which will account for noise and about which more will be said shortly. Once again we assume that the gain has a very long recovery time and thus that it saturates to a steady-state level due to the passage of many ultrashort pulses according to

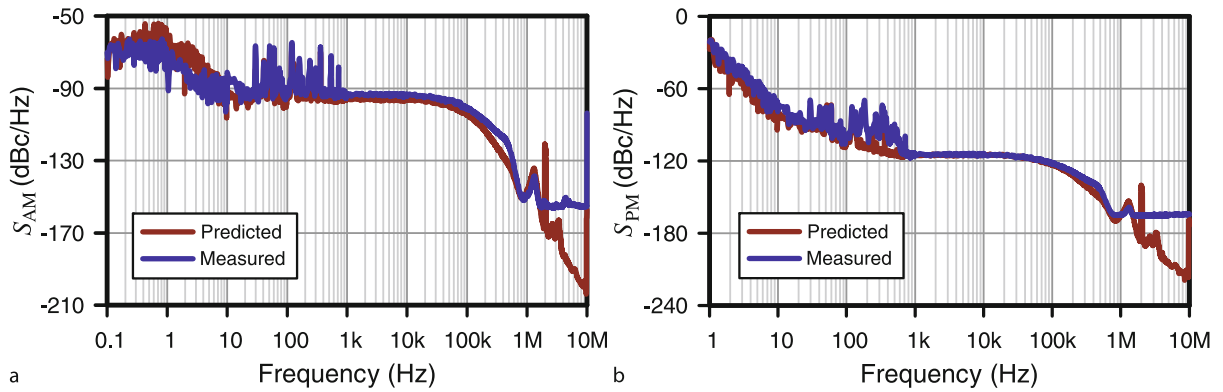
$$g = \frac{g_0}{1 + \frac{1}{P_s T_R} \int |a|^2 dt} \quad (242)$$

where  $P_s$  is the saturation power.



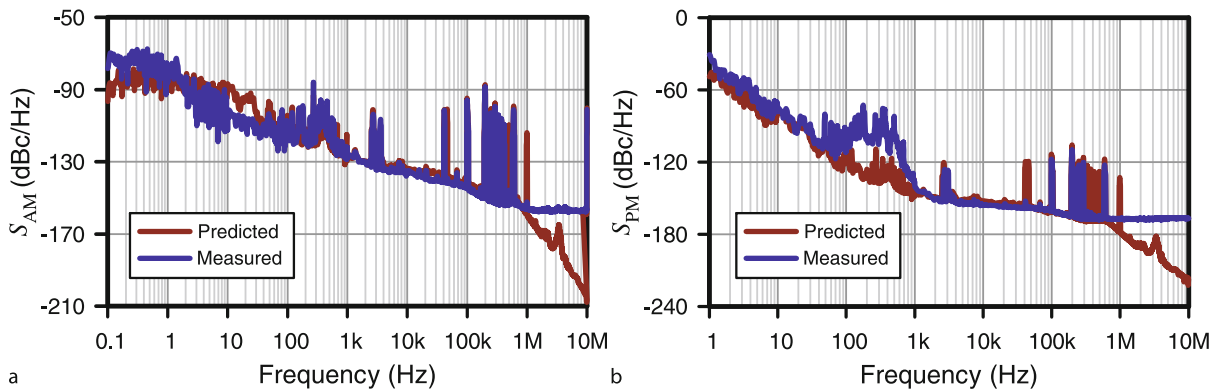
Noise and Stability in Modelocked Soliton Lasers, Figure 33

**a** Pump noise spectra of a single longitudinal-mode diode-pumped solid-state (DPSS) laser and an argon-ion laser. **b** Measured noise transfer functions of KLM Ti:sapphire laser using both pump lasers



Noise and Stability in Modelocked Soliton Lasers, Figure 34

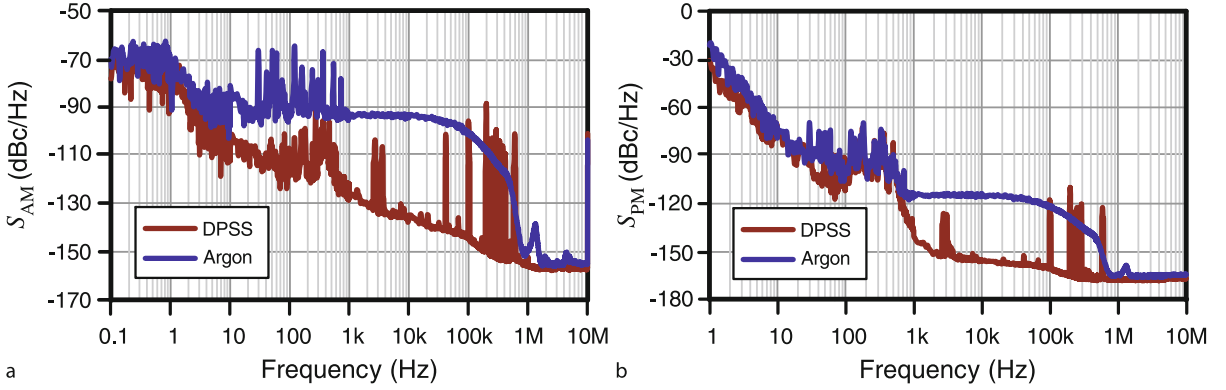
**a** Amplitude noise power spectral density of a modelocked Ti:sapphire laser when pumped by an argon-ion laser running all lines. Compare with the argon-ion laser noise spectrum in Fig. 33a. **b** Single sideband phase noise power spectral density of the same laser



Noise and Stability in Modelocked Soliton Lasers, Figure 35

**a** Amplitude noise power spectral density of a modelocked Ti:sapphire laser when pumped by a DPSS laser. Compare with the argon-ion laser noise spectrum in Fig. 33a. **b** Single sideband phase noise power spectral density of the same laser





**Noise and Stability in Modelocked Soliton Lasers, Figure 36**

**a** Amplitude noise power spectral density of a modelocked Tisapphire laser when pumped by an argon-ion laser running all lines (blue) compared with the same laser pumped by a DPSS pump laser (red). **b** Single sideband phase noise under same conditions as a

Once again we introduce a solution ansatz. Let

$$a(t, T) = A_0 \operatorname{sech} \left[ \frac{1}{\tau} (t - t_0) \right]^{(1+i\beta)} \exp \left( i\psi \frac{T}{T_R} + i\theta \right) \quad (243)$$

which generates the pair of equations

$$-i\psi + g - l + \frac{(1+i\beta)^2}{\tau^2} \left( \frac{g}{\Omega_g^2} + iD \right) = 0 \quad (244)$$

$$\frac{2 + 3i\beta - \beta^2}{\tau^2} \left( \frac{g}{\Omega_g^2} + iD \right) = (\gamma - i\delta) A_0^2 \quad (245)$$

and where

$$t_0 = \frac{gT}{\Omega_g T_R}. \quad (246)$$

From these we can find stable solutions for the physical parameters  $\tau$  (pulsewidth),  $\beta$  (chirp parameter),  $\psi$  (phase shift), and  $g$ , the gain. Of the possible stable solutions, pulses that contain no chirp ( $\beta = 0$ ) are to be preferred since they minimize the time-bandwidth product and thus minimize the pulsewidth for a given gain-bandwidth. With this in mind we set  $\beta = 0$  and find, from (244) that the following relation obtains;

$$\frac{g/\Omega_g^2}{-D} = \frac{\gamma}{\delta} \equiv \mu. \quad (247)$$

The term  $g/\Omega_g^2$  represents the “gain dispersion”, and its ratio to the dispersion parameter ( $-D$ ) is given by the ratio of the saturable absorption to SPM coefficients. Applying  $\beta = 0$  and (247) above to (244) and (245) results in the following,

$$A_0^2 \tau^2 = \frac{2|D|}{\delta} \quad (248)$$

$$\psi = -\frac{|D|}{\tau^2} = -\frac{\delta}{2} A_0^2 \quad (249)$$

$$g - l = -\frac{1}{\tau^2} \frac{g}{\Omega_g^2} = -\frac{\gamma}{2} A_0^2. \quad (250)$$

If these relations are now substituted into the general form of the solution, (243), we have the following steady-state unchirped solution

$$a_0(t, T) = A_0 \operatorname{sech} \left( \frac{t - t_0}{\tau} \right) \exp \left[ -i \left( \frac{\delta}{2} A_0^2 \frac{T}{T_R} + \theta \right) \right] \quad (251)$$

where the pulsewidth

$$\tau = \frac{1}{A_0} \sqrt{\frac{2|D|}{\delta}}. \quad (252)$$

Now, to study the effects of noise on the stability of the pulse solution, we can employ perturbation theory [90]. In this technique, when applied to a general differential equation such as  $\mathcal{L}\{y(x)\}$  where  $\mathcal{L}$  is an arbitrary differential operator, we assume that a small adjustable perturbation,  $\xi$ , can be inserted strategically into the problem such that when  $\xi = 0$ , the problem is exactly solvable. This gives the zero-order solution  $y_0(x)$ . We can then construct a perturbation solution in terms of powers of  $\xi$ ,

$$y(x) = y_0(x) + \xi y_1(x) + \xi^2 y_2(x) + \dots \quad (253)$$

Each of the  $y_n(x)$  are computed from the prior terms  $y_0(x), y_1(x), \dots, y_{n-1}(x)$ .

Soliton perturbation theory takes a slightly different approach [91]. To allow for the possibility of fluctuations

in the pulse parameters owing to noise, we make the new ansatz

$$a(t, T) = A_o [a_s(t - t_o) + \Delta a(t - t_o, T)] \exp \left( -i \frac{\delta}{a} A_o^2 \frac{T}{T_R} \right) \quad (254)$$

where

$$a_s(t) = A_o \operatorname{sech} \left( \frac{t}{\tau} \right). \quad (255)$$

$a_s(t)$  is the perturbation. Substituting the ansatz (254) into the master equation (241) and retaining terms to first order in  $\Delta a$  leads to a new equation for  $\Delta a$ ;

$$\begin{aligned} T_R \frac{\partial}{\partial T} \Delta a(t - t_o, T) &= (\mu - i) \left\{ -\frac{\delta}{2} A_o^2 + |D| \frac{\partial^2}{\partial t^2} + 2\delta a_s^2(t - t_o) \right\} \delta a \\ &\quad + (\mu - i) \delta a_s^2(t - t_o) \Delta a^* \\ &\quad - g_s \left( 1 - \frac{1}{\Omega_g} \frac{\partial}{\partial t} \right) \frac{a_s(t)}{2\tau A_o^2} \int dt a_s(t) (\Delta a + \Delta a^*) \\ &\quad + T_R S(t, T) \end{aligned} \quad (256)$$

where

$$g_s \equiv \frac{g_o}{P_s T_R} \frac{2\tau A_o^2}{\left( 1 + \frac{2\tau A_o^2}{P_s T_R} \right)^2}. \quad (257)$$

The goal now is to expand the perturbation in a series of terms which independently account for fluctuations in the relevant physical parameters. Let these be called the “perturbation amplitudes”  $\Delta w(T)$  (pulse energy),  $\Delta \theta(T)$  (phase),  $\Delta t(T)$  (timing) and  $\Delta p(T)$  (frequency). We now expand the perturbation in terms of these amplitudes and derivatives of the steady-state solution

$$\begin{aligned} \Delta a &= f_w(t - t_o) \Delta w(T) + f_\theta(t - t_o) \Delta \theta(T) \\ &\quad + f_t(t - t_o) \Delta t(T) + f_p(t - t_o) \Delta p(T) + \Delta a_c(t - t_o, T) \end{aligned} \quad (258)$$

where  $\Delta a_c(t - t_o, T)$  accounts for continuum radiation which is not part of the pulse. The sensitivities to the perturbation amplitudes are given by derivatives of the steady-state solution  $a_o$ ,

$$\text{ENERGY} \quad f_w(t) = \frac{\partial a_o}{\partial w_o} = \frac{1}{w_o} \left[ 1 - \frac{t}{\tau} \tanh \left( \frac{t}{\tau} \right) \right] a_s(t) \quad (259)$$

$$\text{PHASE} \quad f_\theta(t) = \frac{\partial a_o}{\partial \theta_o} = i a_s(t) \quad (260)$$

$$\text{TIMING} \quad f_t(t) = \frac{\partial a_o}{\partial t_o} = \frac{1}{\tau} \tanh \left( \frac{t}{\tau} \right) a_s(t) \quad (261)$$

$$\text{FREQUENCY} \quad f_p(t) \equiv i \frac{2}{w_o} t a_s(t). \quad (262)$$

The next step involves defining a set of orthogonal adjoint functions to those (259)–(262) above and then projecting upon them the coefficients (amplitudes) of the perturbation expansion [91].

$$\underline{f}_w(t) = 2a_s(t) \quad (263)$$

$$\underline{f}_\theta(t) = i \frac{2}{w_o} \left[ 1 - \frac{t}{\tau} \tanh \left( \frac{t}{\tau} \right) \right] a_s(t) \quad (264)$$

$$\underline{f}_t(t) = \frac{2}{w_o} t a_s(t) \quad (265)$$

$$\underline{f}_p(t) = i \left[ \frac{2}{w_o \tau} \tanh \left( \frac{t}{\tau} \right) \right] a_s(t). \quad (266)$$

These functions are chosen to be orthonormal so that

$$\operatorname{Re} \left\{ \int \underline{f}_i^*(t) f_j(t) dt \right\} = \delta_{ij}. \quad (267)$$

The projections of the perturbation amplitudes are then

$$\Delta w(T) = \frac{1}{2} \int \left[ \underline{f}_w^*(t) \Delta a(t) + \underline{f}_w(t) \Delta a^*(t) \right] dt \quad (268)$$

$$\Delta \theta(T) = \frac{1}{2} \int \left[ \underline{f}_\theta^*(t) \Delta a(t) + \underline{f}_\theta(t) \Delta a^*(t) \right] dt \quad (269)$$

$$\Delta p(T) = \frac{1}{2} \int \left[ \underline{f}_p^*(t) \Delta a(t) + \underline{f}_p(t) \Delta a^*(t) \right] dt \quad (270)$$

$$\Delta t(T) = \frac{1}{2} \int \left[ \underline{f}_t^*(t) \Delta a(t) + \underline{f}_t(t) \Delta a^*(t) \right] dt. \quad (271)$$

When the perturbation expansion (258) using the projections (268)–(271) is substituted back into the differential equation for the original perturbation ansatz (254), we obtain the equations of motion for the perturbation amplitudes

$$T_R \frac{\partial}{\partial T} \Delta w = (-2g_s + 2\gamma A_o^2) \Delta w + T_R S_w(T) \quad (272)$$

$$T_R \frac{\partial}{\partial T} \Delta \theta = -\delta A_o^2 \frac{\Delta w}{w_o} + T_R S_\theta(T) \quad (273)$$

$$T_R \frac{\partial}{\partial T} \Delta p = -\frac{4}{3} \frac{g}{\Omega_g^2 \tau^2} \Delta p + T_R S_p(T) \quad (274)$$

$$T_R \frac{\partial}{\partial T} \Delta t = -2|D| \Delta p - \frac{g}{\Omega_g} \frac{\Delta w}{w_o} + T_R S_t(T) \quad (275)$$

where the  $S_j(T)$  are the noise sources defined in terms of their projections

$$S_j(T) = \frac{1}{2} \int \left[ \underline{f}_t^*(t) S(t, T) + \underline{f}_t(t) S^*(t, T) \right] dt. \quad (276)$$

The physical origins of the  $S(t, T)$  will be described shortly.

The equations of motion (272)–(275) are all linear first-order inhomogeneous equations driven by noise terms. Equations (272) and (274) are simple relaxation equations while the other two have no natural relaxation times and are driven by other perturbation terms as well as the noise. What can we learn about the general nature of these equations? The general form of a driven relaxation equation is

$$\frac{dy(t)}{dt} + \alpha y(t) = F(t) \quad (277)$$

where  $F(t)$  is a forcing function. In the context of an initial-value problem, the solution is found by using an integrating factor [92],

$$y(t) = y(0)e^{-\alpha t} + \int_0^t e^{-\alpha(t-t')} F(t') dt'. \quad (278)$$

After complete relaxation of the initial condition ( $t \gg 1/\alpha$ ), the response is given by the convolution of the forcing function with the decay term. Comparing (277) with (272)–(275), we see that the noise terms represent the forcing function. Equation (272) says that if the saturable absorption action is strong enough (large  $\gamma_o A_o^2$ ), the pulse energy will grow exponentially unless the gain saturation ( $g_s$ ) is strong enough to suppress it. Equation (273) indicates that pulse energy fluctuations induce phase shift due to the Kerr effect ( $\delta$ ). Equation (274) says that if the center frequency of the pulse is tuned off line center, it will be pushed back ( $\Delta p \rightarrow 0$ ) because the gain is highest at line center. In the last equation, (275), we see that both frequency change and pulse energy change cause timing shift. The effect of the first term because of group-velocity dispersion ( $|D|$ ) and the effect of the second term because of gain changes producing index of refraction change due to the Kramers–Kronig relations.

### Physical Sources of Noise

There are several principal sources of noise that will enter into the equations of motion (272)–(275) via the projections (276). The first is due to gain fluctuation. This will come about most simply by virtue of a noisy pump mechanism. From the master equation (241), we see that gain

fluctuations that occur on a relatively long time scale can be described by the equivalent noise source

$$S_{\Delta g}(t, T) = \frac{\Delta g(T)}{T_R} \left( 1 - \frac{1}{\Omega_g} \frac{\partial}{\partial t} \right) a_s(t, T). \quad (279)$$

There is another, slightly subtle, effect that occurs due to gain fluctuation. If the gain and loss curves due not share the exact same line center, then as the gain goes up and down, the *net gain* (that is, the gain-loss) will necessarily shift, and this will cause a frequency-pulling term

$$S_{\Delta g}(t, T) = i \frac{\Delta g(T)}{4T_R(\omega_o - \omega_{oo})} \frac{d}{dt} a_s(t, T) \quad (280)$$

where  $\omega_o$  is the peak frequency of the net gain and  $\omega_{oo}$  is the peak frequency of the gain medium.

The next physical effect is that of cavity length variation and the related effect of index fluctuations. The effective noise source is given by

$$S_{\Delta L}(t, T) = -\frac{\Delta L(T)}{T_R} \left( i \frac{\omega_o}{c} n + \frac{1}{v_g} \frac{d}{dt} \right) a_s(t, T). \quad (281)$$

The first term describes frequency pulling while the second term describes timing jitter. If we also add the effect of index variations,  $\Delta n$  then we simply replace  $\Delta L$  by  $L(\Delta n/n)$ .

The final physical effect we take into account is that of spontaneous emission from the gain medium into the cavity. Spontaneous emission has a white spectral density with a correlation function

$$\langle S_{qn}(T, t) S_{qn}^*(T', t') \rangle = \xi \frac{2g}{T_R} h\nu \delta(T - T') \delta(t - t') \quad (282)$$

where  $\xi$  is an enhancement factor that accounts for incomplete inversion of the atomic gain medium.

With the physical sources  $S_{\Delta i}$  defined (279)–(282), we can project them onto the adjoint functions using the definition (276) to obtain the noise sources

$$S_w(T) = 2 \frac{\Delta g(T)}{T_R} w_o + S_{w,qn}(T) \quad (283)$$

$$S_\theta(T) = \frac{\omega_o}{v_g T_R} \left[ \Delta L(T) + L \frac{\Delta n(T)}{n} \right] + S_{\theta,qn}(T) \quad (284)$$

$$S_p(T) = \frac{\Delta g(T)}{3T_R(\omega_o - \omega_{oo})\tau^2} + S_{p,qn}(T) \quad (285)$$

$$S_t(T) = \frac{\Delta g(T)}{T_R \Omega_g} \frac{1}{T_R v_g} \left[ \Delta L(t) + L \frac{\Delta n(T)}{n} \right] + S_{t,qn}(T) \quad (286)$$

where the  $S_{i,qn}(T)$  are the quantum noise sources with the correlation functions

$$\langle S_{i,qn}(T) S_{h,qn}(T') \rangle = D_{i,qn} \delta_{i,h} \delta(T - T') \quad (287)$$

and the following related diffusion coefficients

$$D_{w,qn} = 4w_0 \xi \frac{2g}{T_R} h\nu \quad (288)$$

$$D_{\theta,qn} = \frac{4}{3w_0} \left(1 + \frac{\pi^2}{12}\right) \xi \frac{2g}{T_R} h\nu \quad (289)$$

$$D_{p,qn} = \frac{8}{3w_0^2 \tau^2} \xi \frac{2g}{T_R} h\nu \quad (290)$$

$$D_{t,qn} = \frac{\pi^2 \tau^2}{3w_0} \xi \frac{2g}{T_R} h\nu. \quad (291)$$

We can now include (283)–(286) as the noise source terms in the equations of motion of the perturbation expansion (272)–(275).

At this point, (272)–(275) constitute a set of deterministic equations of motion for the coefficients,  $\Delta w$ ,  $\Delta \theta$ ,  $\Delta p$  and  $\Delta t$  of the perturbation expansion for  $\Delta a$  (258). If the noise sources were deterministic functions, we could solve these directly but, noise sources are random processes, and thus we must resort to statistical analyses using correlation and spectral density functions. This would take us down the correct path but would be a rather complicated addition to our already long story. Suffice it to say that Haus and Mecozzi have considered the implications of this by making reasonable assumptions about the spectral densities of the noise sources and have been able to predict the shapes of the spectra of the physical observables [65].

### Large Scale Instability in Soliton Lasers

The soliton concept as applied to passively-modelocked lasers came about rather late in the general development of modelocking but was crucial to moving forward the understanding and engineering of this very important technology. Before discussing large-scale instability in soliton-like lasers, it is useful to take a brief look at instability in the other class of modelocking known as “active modelocking” [93,94]. In this approach, the net gain of the laser is modulated at the exact round-trip transit time within the cavity. This preferentially amplifies a section of the circulating fields which evolve into a pulse with pulsewidth on the order of (or better than) the width of the pumping pulse (within the limits set by the bandwidth of the gain medium). The gain itself can be directly modulated by, for example, a time varying pump, or a variable loss element such as an electro- or acousto-optic modulator inserted in the cavity. Another approach is to place an FM modulator in the cavity. A short pulse becomes synchronized with the peak of the modulation since the sides of the pulse receive strong FM modulation and the energy is shed outside of the bandwidth of the gain medium. From the standard

theory of active modelocking [93], the pulsewidth is proportional to

$$\tau_p \propto \left(\frac{1}{\Delta_m}\right)^{1/4} \left(\frac{1}{f_m}\right)^{1/2} \quad (292)$$

where  $\Delta_m$  is the modulation index (see Subjects. “Analytical Description of Envelope Noise”–“Fourier Analysis of Sinusoidal Phase Modulation”) and  $f_m$  is the modulation frequency. The modulation index varies with drive power  $P_m$  as

$$\Delta_m \propto \begin{cases} P_m, & \text{acousto-optic AM modelockers,} \\ P_m^{1/2}, & \text{electro-optic FM modelockers.} \end{cases} \quad (293)$$

Substituting these dependencies into (292) we find

$$\tau_p \propto \begin{cases} P_m^{-1/4} f_m^{-1/2} & \text{acousto-optic AM modelockers,} \\ P_m^{-1/8} f_m^{-1/2} & \text{electro-optic FM modelockers.} \end{cases} \quad (294)$$

We see that there is a greater advantage to operating at higher frequencies but the frequency is related to the round-trip cavity time  $2L/c$  so they must be considered together. On the other hand, one can operate the modelocker at harmonics of the fundamental cavity rate and this succeeds at producing shorter pulses but with lower peak power since they also take on the same repetition rate as the modelocker frequency.

Early reports of instability in actively modelocked lasers included theoretical studies of pump laser and modelocked laser cavity length mismatch [95] and experimental studies showing period multiplication and division in a synchronously-pumped modelocked dye laser [96]. A synchronously-pumped modelocked laser is one which is pumped by another modelocked laser which has its cavity round-trip time (nominally) equal to the that of the pumped laser. This is a particular example of the technique of active modelocking described above. The authors varied both the pump laser intensity and the ratio of the pump laser cavity length to the modelocked laser cavity length. The interesting result from this investigation was the following; if the ratio of the pumping laser cavity length  $L$  to the length of the dye laser cavity  $L$  was a rational number,  $m/n$ , then the dye laser would produce a stable train of pulses at a rate of  $m/p$  times the pumping rate, where  $p$  is an integer which depends on the magnitude of the pump power. At high pump power,  $p = 1$ , and at low pump power,  $p \rightarrow n$ . This result is intuitively clear in the sense that the pump pulse defines the maximum gain point in time. If you plot the pulses with the period ratio as

a rational number, then the temporal overlap of the pulses gives the basic timing of the dye laser while if the gain is high enough, pulses in between will also form. In addition to frequency multiplication, period multiplication was also seen to occur. The transitions between stable operating regimes could be considered as bifurcations and they often were accompanied by strong chaotic behavior just at the boundaries. Also, by making small ( $\Delta L/L \approx 2 \times 10^{-4}$ ) adjustments in the dye laser cavity length, they found that the modelocking became highly irregular and chaotic.

MacFarlane and Casperson also found cavity length-dependent instability in an acousto-optically modelocked argon-ion laser [97]. Here, an acousto-optic amplitude modulator is placed within the cavity to act as a time-varying loss mechanism. An acoustic standing-wave field is established at half the round-trip cavity frequency so that every half cycle the fields collapse to zero and allow maximum optical transmission. This establishes the minimum loss point in the temporal dynamics of the laser system. They found that by detuning the acousto-optic modelocker frequency by as little as one part in  $10^4$ , instability set in and further increasing the detuning resulted in strong spectral components at one fourth of the normal operating repetition rate (period quadrupling). Detuning by as much as  $1.6 \times 10^{-4}$  caused a new feature: very low frequency modulation (600 KHz) of the 81 MHz pulse train. Curiously, the trends in instability were not symmetric with respect to detuning. The strong period-four component was only present for negative relative detuning, whereas positive detuning caused an immediate change to chaotic pulsing and overall degradation of laser output. Scavennec, in a theoretical study [95], also found asymmetries with detuning. It is worth noting that the relative frequency detuning is of the same order as the relative cavity mismatch that led to chaos in [96]. Further interesting measurements in the time and radiofrequency domain were also carried out later by MacFarlane et al., [98].

Another type of instability can be seen when multiple modelockers are placed within an optical cavity to achieve shorter modelocked pulses than would otherwise be produced by one alone. Here we have an additional degree of freedom which is the relative phase of the signals driving the modulators. Scott et al., [99], placed an AM and FM modulator within a Nd:YAG laser cavity and drove the AM (acousto-optic) modulator at 40 MHz to produce fundamental modelocking at 80 MHz and the FM (electro-optic) modulator at 1.76 GHz to produce a chirp on the pulse. This combination of fundamental and high-harmonic FM modelocking achieved a pulsewidth reduction of 5.3 times over operating with the AM modelocker alone. When the relative phase between the RF drive signals to

the two modelockers was adjusted, the output pulse was seen to move within the peak of the AM transmission window up to the point where there was insufficient gain to compensate for the intracavity losses. At this point, instability occurred as a new pulse tried to emerge at the maximum gain point, but with opposite chirp. The oppositely chirped pulse was less stable and had lower peak power. Further phase shift brought this pulse to the boundary of the gain > loss region and an unstable transition to a new pulse at the beginning of the gain > loss region occurred.

A large shift in the ultrafast laser community occurred around 1985 with the development of the colliding pulse modelocked dye laser incorporating intracavity dispersion compensation and adjustable saturable absorption [100]. It was recognized that there was an important interplay between the group velocity dispersion, the saturable gain and the saturable absorption and that by controlling them, optimally short pulses (< 27 fs) could be obtained. Shortly thereafter, issues of large scale instability began to receive attention and Avramopoulos et al., [101] presented theoretical and experimental results showing a causal link between certain values of intracavity dispersion and self-phase modulation and overall laser instability. The large scale amplitude fluctuations that were predicted and seen had modulation periods on the order of 100  $\mu$ s which is many hundreds of round-trip cavity times.

The refinement of the twin jet colliding-pulse modelocked dye laser really set the stage for the solid state replacement a few years later where the gain medium was the Ti:sapphire crystal (with its extraordinary bandwidth) and the nonlinearity responsible for self-phase modulation and effective saturable absorption was the Kerr effect in the gain crystal itself [102]. Thus began the era of the all solid-state modelocked soliton laser. Analysis of how these lasers work proceeded rapidly and many approaches were developed, such as the master equation of Haus [64]. Later, workers also expanded the analysis to include spatial effects within the cavity [103,104,105,106]. We will not go into these techniques here since we wish to concentrate of issues of large scale instabilities.

The analytical development of the preceding Sect. “Noise in Soliton Lasers” focused on the impact of small-scale perturbations to the equations of motion for the soliton laser. To study the larger scale phenomena of instability, even more sophisticated techniques have been developed. For example, Tartwijk and Agrawal developed a modal analysis which expanded the electric field and polarization in a fiber laser in a series of the longitudinal modes and then applied the expansion to the Maxwell-Bloch equations [23]. This results in a family of Lorenz-type equations for each of the longitudinal mode expan-



sion coefficients and includes the effects of self-phase modulation, cross-phase modulation and four-wave mixing. Calculations were carried out for single mode operation but, unfortunately, they did not study the case of many modes required to simulate modelocked operation. (We conjecture that the computational resources required to accurately simulate hundreds or thousands of modes over long time scales would be formidable).

Sergeev et al., [107], studied soliton-like solutions of the Ginzburg–Landau equation (GLE, similar in form to the nonlinear Schrödinger equation (90)) which includes saturable absorption, group-velocity dispersion, finite gain-bandwidth and a Kerr nonlinearity. This equation has a hyperbolic-secant solution very similar to the master equation ansatz of Haus and Mecozzi (243). The proceeded to apply the “aberrationless approximation” which involves substituting the sech solution back into the GLE, assume the parameters for pulse amplitude, duration and chirp are functions of the propagation distance (cavity round trips,  $z$ ) and then expand these expressions in a Taylor series to obtain new first-order differential equations for the parameters. Stationary equilibria are sought as a function of the saturation energy of the gain medium. By studying the solutions in the phase plane, they were able to determine an analytical criterion for stable soliton solutions in the weak absorber saturation approximation.

In 1993, Liu and Prucnal [108] discovered a new source of instability in the modelocked soliton laser. A key feature of the early generation of these lasers was the incorporation of an intracavity aperture, or slit, which preferentially passed a tightly focused beam and therefore was lossy for larger diameter beams. This forces the laser to self-modelock by virtue of the Kerr-lens effect which only produces lensing when a high peak-power pulse passes through the gain crystal. They found that by closing down the slit about 10% smaller than that for optimum pulsewidth, large amplitude modulation was obtained with depths approaching 100% and a modulation period of several  $\mu$ s. The process can be understood qualitatively as follows: when the slit is closed down past the size that produces the best pulses, the intracavity loss is increased because of the aperturing effect on the beam. Thus, the intracavity circulating power goes down and there is less stimulated emission in the gain medium. Less stimulated emission allows the population inversion to increase beyond its normal steady-state value which means the gain increases faster than the cavity loss rate and overshoots the steady-state value which would normally balance the static cavity losses. Now the fields start building exponentially fast which increases the Kerr effect and the self-focusing further reducing the cavity losses since the beam is now focused tighter through

the narrowed aperture. At this point, two things happen. First, in overshooting the steady-state modelocked pulse amplitude, it generates a focus that is actually too small for the aperture in the sense that the divergence (convergence) angle increases and the focus shifts, which again starts to increase loss. Second, the higher field amplitude starts to eat up the population inversion by stimulated emission which also contributes to the downward trend of the field amplitude and thus the cycle repeats. The fluctuation period (or frequency) depends on the usual time constants (cavity and energy decay times) as well as the pumping rate,  $r$ . Just as in the relaxation oscillation analysis (Subsect. “Pump-Induced Noise and the Noise Transfer Function”) the prime frequency of the induced modulation is very nearly the relaxation oscillation frequency (199).

If one couples an external cavity with a nonlinear element such as a single-mode fiber to the main laser cavity, the pulses in each will combine interferometrically and shorten owing to the nonlinear phase shift in the wings of the external cavity pulse. This is called “additive pulse modelocking” and was quite popular for a number of years. Because of the presence of the nonlinearity and the time-delayed feedback into the main laser cavity, this system is a perfect candidate for complex dynamics. Indeed, Sucha et al., [109] observed period doubling and quasi-periodic modelocking in a NaCl F-center laser. They also numerically simulated the laser system and found that as the fiber nonlinearity was increased from zero, the laser starts out in cw operation and then proceeds to stable modelocking. Beyond this point a bifurcation in the pulse energy takes place and, thus, period doubling. Further increases in the nonlinearity results in more period doubling and then an abrupt transition to chaos. In the period doubling region they found that the main cavity and external fiber cavity exchanged pulses on alternate round trips and these pulse had differing temporal and spectral character.

As ideas from chaos theory migrated into the laser community in the early days with studies on CW lasers, we see that they eventually found their way into analyses on modelocked lasers. The bifurcation and period doubling phenomena described above are good examples as well as the analysis by Sánchez and Hnilo who used Poincaré (iterative) maps in the complex plane to describe Kerr-lens modelocked soliton lasers [110]. We wish, now, to conclude this chapter with a brief discussion of what is happening in this field in recent years, because the emphasis has shifted back to issues of small-scale perturbation. In particular, the advent of the carrier-envelope stabilized, octave-spanning modelocked soliton laser has opened a new door onto the world of precision timekeeping and metrology and the fundamental mechanisms gov-

erning timing stability and laser linewidth are once again extremely important topics.

### Laser Clocks

Researchers in the ultrafast optics community have worked tirelessly to make their laser pulses ever-shorter and evermore stable since the invention of the modelocking technique. Short pulses with a high degree of pulse-to-pulse stability are a powerful tool for time-domain studies of ultrafast phenomena in physics, chemistry, biology and electronics. But we can also appreciate this from a frequency-domain viewpoint. An ideal train of ultrashort laser pulses has an associated frequency spectrum composed of narrow, equally-spaced “comb-lines” with spacing equal to the inverse of the time between the pulses (see Fig. 16). If the exact frequency of one of the comb lines is known, and the repetition rate (interline spacing) is accurately known, then we have in effect a “ruler” in the frequency domain which can serve as a standard for measuring other optical sources or spectroscopic features. Recently, a revolution occurred in the field of laser physics when two teams reported a method for stabilizing the repetition rate of a modelocked laser using the “self-referencing” principle based on an  $f - 2f$  interferometer and an octave-spanning modelocked laser source [111,112]. This invention has heralded a new era in precision time and frequency metrology and new laser-based clockworks are now being created all over the world. Indeed, two out of three recipients of the 2005 Nobel Prize in Physics were recognized for their contributions to this invention and the significance of the achievement [113]. These laser systems have now demonstrated more than three orders of magnitude better timing stability than the hyperfine transition of the cesium atom which forms the basis of the currently accepted definition of the second. They are also demonstrating one to two orders of magnitude better stability than some of the other more exotic atomic standards such as the hydrogen maser and some trapped ions. In essence, the frontiers of time and frequency metrology have just advanced by more than two orders of magnitude in the short span of several years. (For good reviews on the evolution of this laser technique and precision spectroscopy in general, see [19,111,114] and references therein).

### Modelocked Laser Comb Lines and Frequency Metrology

A key concept in understanding the aforementioned revolution in modelocked lasers is the “carrier-envelope offset” in the pulses that exit the laser cavity. Optical pulses

can be decomposed into their carrier and their envelopes. The carrier propagates at the phase velocity in a medium determined by the index of refraction,  $v_p = c/n(\omega)$ . The envelope propagates at the group velocity  $v_g = d\omega/d\beta$ . Thus, for a small propagation distance  $\delta z$ , there will be a phase slip between the center of mass of the envelope and the carrier given by

$$\delta\phi_{\text{CEO}} = \left( \frac{1}{v_g} - \frac{1}{v_p} \right) \omega_c \delta z \quad (295)$$

where the subscript “CEO” stands for “carrier-envelope offset”. If we integrate this function along the whole path of the laser cavity, we arrive at a total round-trip carrier-envelope phase offset

$$\Delta\phi_{\text{CEO}} = \left( \frac{1}{v_{g\text{avg}}} - \frac{1}{v_{p\text{avg}}} \right) \omega_c p_c \quad \text{modulo } 2\pi \quad (296)$$

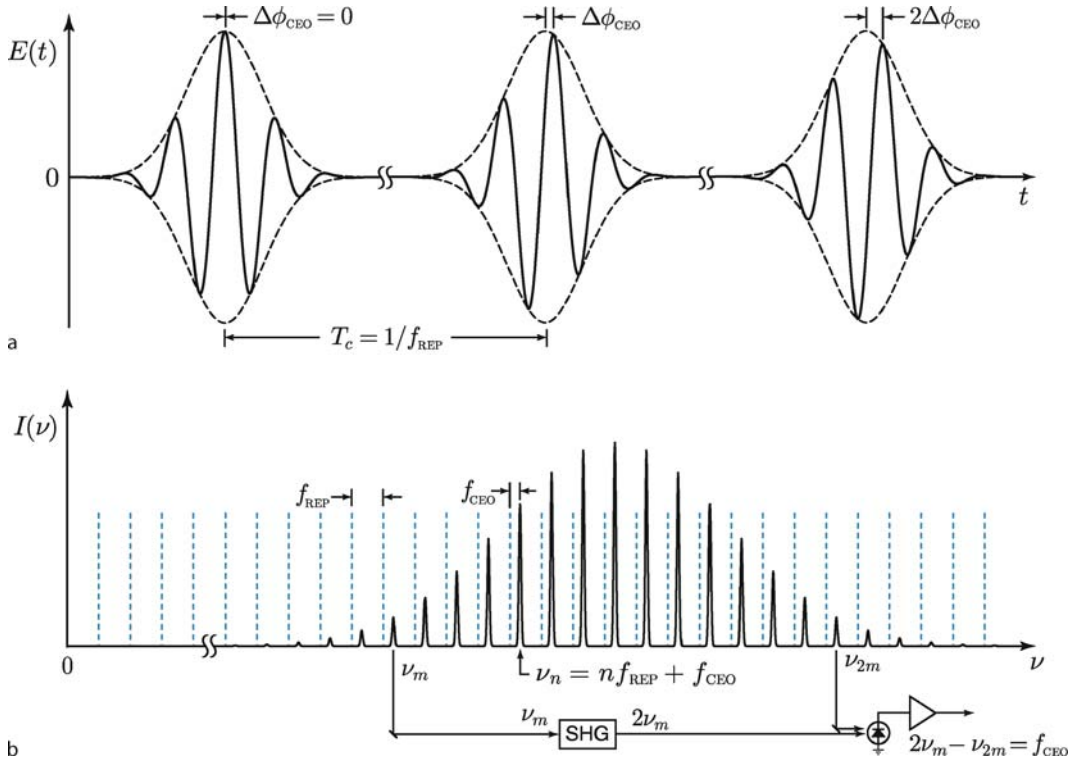
where  $v_{g\text{avg}}$  and  $v_{p\text{avg}}$  are the path-averaged group and phase velocities and  $p_c$  is the total round-trip cavity path length. We include in our definition the modulo  $2\pi$  since the total phase shift can be several hundred optical cycles but the electric field is  $2\pi$  periodic and thus the net number of cycles is not seen [115]. The frequency of the recurring carrier-envelope phase offset is given by

$$f_{\text{CEO}} = \frac{1}{2\pi} \frac{d\phi_{\text{CEO}}}{dt} = \frac{1}{2\pi} \frac{\Delta\phi_{\text{CEO}}}{T_c} = \frac{\Delta\phi_{\text{CEO}}}{2\pi} f_{\text{REP}} \quad (297)$$

where  $f_{\text{REP}}$  is the repetition rate of the envelope of the modelocked pulse train. The relationship between the carrier and the envelope timing is shown in Fig. 37a. In Fig. 37b we see how the frequencies are related to each other. The dashed lines show harmonics of the fundamental repetition rate while the solid lines are the actual comb spectrum of the laser. The frequency of any comb line is given by  $\nu_n = n f_{\text{REP}} + f_{\text{CEO}}$ .

There are two requirements for using this frequency comb as a high precision basis for optical frequency measurements: 1. the comb has to be stabilized against fluctuations in  $f_{\text{REP}}$  and 2.  $f_{\text{CEO}}$  must be determined. Then by using a coarse frequency measuring instrument (such as a spectrometer or wavemeter) one can correctly find the integer  $n$  and all that needs to be done is to accurately determine  $f_{\text{CEO}}$  to have a very precise optical frequency measuring instrument.

The method of stabilizing  $f_{\text{REP}}$  by using an octave-spanning spectrum and cavity feedback control by self-referencing is exceptionally clever [111]. First, we must prepare a spectrum of comb lines that spans an octave. This was originally done using a highly nonlinear fiber as



**Noise and Stability in Modelocked Soliton Lasers, Figure 37**

**a** Time domain representation of the laser pulse electric field showing the carrier-envelope phase offset  $\Delta\phi_{\text{CEO}}$ . **b** Laser power spectrum showing comb lines separated by the cavity repetition frequency  $f_{\text{REP}}$ . The absolute frequencies of the comb lines are shifted from integer multiples of  $f_{\text{REP}}$  by the carrier-envelope offset frequency  $f_{\text{CEO}}$ . Also shown is the  $f - 2f$  interferometer for absolute frequency stabilization which is based on the premise that the spectrum spans at least one octave so that a low frequency component  $\nu_m$  can be frequency doubled and beat against a nearby comb line at  $\nu_{2m}$

which caused dramatic spectral broadening due to self-phase modulation. In fact, the development of specialty microstructure hollow-core or “photonic crystal” fiber was the key to the birth of the new technique (see [116,117] and references therein). Suppose we spread out the spectrum with a diffraction grating, pick off a comb line at the low frequency (infrared) side,  $\nu_m$  and frequency double it with a second harmonic generation crystal (SHG). Simultaneously, we pick a comb line at the high frequency side of the spectrum, which is exactly twice the index number,  $2m$ , of the low frequency comb. We then combine the light and shine it on a photodiode. The current that results is a sinusoid at the beat or difference frequency between the two. That is,

$$\begin{aligned}\Delta f &= 2\nu_m - \nu_{2m} \\ &= 2(mf_{\text{REP}} - f_{\text{CEO}}) - (2mf_{\text{REP}} - f_{\text{CEO}}) = f_{\text{CEO}}.\end{aligned}\quad (298)$$

This provides us with a direct measure of the carrier offset frequency and thus knowledge of the actual frequency

of any comb line. More importantly, that difference frequency,  $f_{\text{CEO}}$  can be used as an error signal to control the cavity length and the net difference in group and phase velocities and thus drive  $f_{\text{CEO}} \rightarrow 0$ . This is the “self-referencing” that has revolutionized the stability of the modelocked lasers and transformed the field of time and frequency metrology [17,18,19,111,114]. As can be seen in the explosion of literature on the subject in recent years, there are many new applications for this technique, both for science and engineering, and thus questions about ultimate stability, linewidth, noise, etc., are currently a hot area of research. We will now briefly touch on a few of the most recent results.

It was recognized fairly early on that noise accompanying the pump source could be a significant source of noise in the Kerr-lens modelocked lasers (as well as any other laser) [81,82,83,84,85]. As discussed in Subsect. “Pump-Induced Noise and the Noise Transfer Function” of this chapter, the principal mechanisms for coupling pump amplitude fluctuations to modelocked laser noise are population inversion (gain) modulation for am-

plitude noise and 1. direct index modulation, 2. thermal modulation, 3. nonlinear effects, and 4. beam-pointing instability due to thermal and direct index fluctuations for the phase noise. Scott et al., [85], have been primarily concerned with the effect on the modelocked laser envelope while Xu et al., [118] Helbing et al., [81] and Matos et al., [119] studied the effect on the carrier-envelope offset frequency. The former is important for free-running (non-locked) lasers while the latter is important for the self-referenced and locked Kerr-lens lasers. The frequency-dependence of the coupling mechanisms is described in detail in Subsect. “**Pump-Induced Noise and the Noise Transfer Function**”. We also note here that, since the low frequency phase sensitivity increases with pump power as the frequency is lowered (to DC), the pump power can be used to effectively control the modelocking frequency [120].

A detailed study of the variation in absolute and relative comb frequencies was carried out by Holman et al., [121]. They derived expressions for the sensitivity of  $f_{\text{REP}}$  and  $f_{\text{CEO}}$  to the peak circulating pulse intensity,  $I$  by differentiating

$$\frac{df_{\text{REP}}}{dI} = \frac{1}{L_c} \frac{dv_g}{dI} \quad (299)$$

$$\begin{aligned} \frac{df_{\text{CEO}}}{dI} = & \frac{1}{2\pi} \frac{\partial \omega_c}{\partial I} \left( 1 - \frac{v_g}{v_p} \right) \\ & + \frac{\omega_c}{2\pi} \frac{v_g}{v_p} \left( \frac{1}{v_p} \frac{dv_p}{dI} - \frac{1}{v_g} \frac{dv_g}{dI} \right) \end{aligned} \quad (300)$$

where  $\omega_c$  is the spectrally-weighted center frequency and thus  $\partial \omega_c / \partial I$  is the intensity-driven shift in the optical spectrum. The authors treat only the effect of the variation of the indices of refraction with pulse intensity by letting  $\bar{n} \equiv \bar{n}_0 + \bar{n}_2 I$  be the average refractive index within the laser cavity such that in one round trip,  $\bar{n} L_c$  gives the true optical path length of air plus Ti:sapphire crystal. With this definition, we can now define an effective phase and group velocity,

$$v_p = c/\bar{n}, \quad v_g = c/\left[\bar{n} + \omega_c(d\bar{n}/d\omega)_c\right]. \quad (301)$$

Now, carrying out the indicated differentiations in (299) and (300), we find

$$\begin{aligned} \frac{df_{\text{REP}}}{dI} = & -\frac{1}{L_c} \frac{v_g^2}{c} \left[ \bar{n}_2 + \omega_c \left( \frac{d\bar{n}_2}{d\omega} \right)_{\omega_c} + c \frac{\partial \omega_c}{\partial I} \frac{\partial}{\partial \omega_c} \left( \frac{1}{v_g} \right) \right] \\ & - \frac{\omega_c v_g}{c} \frac{\partial \bar{n}}{\partial \omega_c} \end{aligned} \quad (302)$$

$$\begin{aligned} \frac{df_{\text{CEO}}}{dI} = & \frac{\omega_c^2}{2\pi} \frac{v_g^2}{c^2} \left[ \bar{n}_0 \left( \frac{d\bar{n}_2}{d\omega} \right)_{\omega_c} - \bar{n}_2 \left( \frac{d\bar{n}_0}{d\omega} \right)_{\omega_c} \right] \\ & + \frac{1}{2\pi} \frac{\partial \omega_c}{\partial I} \left[ \left( 1 - \frac{v_g}{v_p} \right) + \frac{\omega_c v_g^2}{v_p} \frac{\partial}{\partial \omega_c} \left( \frac{1}{v_g} \right) \right. \\ & \left. - \frac{\omega_c v_g}{c} \frac{\partial \bar{n}}{\partial \omega_c} \right]. \end{aligned} \quad (303)$$

All of the terms in (302) and (303) are available in the literature except  $\partial \omega_c / \partial I$  and  $\partial v_g^{-1} / \partial \omega_c$ . However, two sets of experimental data can uniquely determine these quantities. In their experiments, the authors placed an acousto-optic amplitude modulator (AOM) between the pump laser and a pair of KLM Ti:sapphire lasers; one configured for 100 MHz and one configured for 750 MHz. For slow pump variations, standard frequency counters were used to monitor  $f_{\text{REP}}$  and  $f_{\text{CEO}}$ . For higher frequency sinusoidal pump modulations, lockin receivers were used (i. e. homodyne receivers with local oscillators derived from the AOM drive signal). For static changes in the pump power, a local minimum in the center frequency,  $\omega_c$  of the pulse spectrum was measured. That is,  $\partial \omega_c / \partial I$  undergoes a sign change with  $I$ . This would imply a possible related zero in  $df_{\text{CEO}}/dI$  (see Eq. 303) above) and indeed, this was confirmed in the data. Possible explanations for the shift in the spectrum with pump power include a possible offset between the gain and the loss bands within the cavity. As the pump power is changed, the gain goes up and down but the losses stay fixed, The equilibrium point for lasing thus becomes frequency-dependent, if the two curves do not exactly overlap. However, this does not provide an explanation for the local minimum. It is interesting, though, and the authors rightly point out that there are strong implications for choosing design parameters in the Ti:sapphire laser to try and minimize the sensitivity to pump fluctuations.

The authors followed the DC characterizations of  $\Delta f_{\text{REP}}$  and  $\Delta f_{\text{CEO}}$  with dynamic measurements by driving the AOM from 10 Hz to 400 KHz. Unlike the noise transfer function measurements presented in Subsect. “**Pump-Induced Noise and the Noise Transfer Function**”, there was not a significant variation with frequency until the very lowest frequency portion of the spectrum (approximately 1 KHz), which they attribute to thermal effects. Once again, there is a definite sign dependence on the average pump power (and hence the circulating peak power). For average output powers below 750 mW, the slope  $\Delta f_{\text{CEO}}/\Delta I > 0$  whereas for output powers above 750 mW,  $\Delta f_{\text{CEO}}/\Delta I < 0$ , which was consistent with the DC data.

Another very interesting result from this work is a study of the nature of the lineshape of the CEO signal. The authors studied this with a radiofrequency spectrum analyzer and found that the linewidth was narrowest at the zero-crossing point of  $\Delta f_{\text{CEO}}/\Delta I$  when the AOM was driven at 10 KHz. The correlation between the minimum linewidth and a zero-crossing for 10 KHz modulation is simply explained by noting that the zero-crossing implies minimum sensitivity to pump fluctuations for frequencies above the thermal response. The paper concludes with a study of the CEO signal linewidth as measured on an RF spectrum analyzer as a function of laser bandwidth. Unfortunately, this type of measurement is not conclusive as the signal-to-noise ratio and the simple technique did not afford unambiguous results. A better method would have been to separate the AM from the PM noise, measuring the AM at baseband and the PM about the  $f_{\text{CEO}}$  carrier using a quadrature mixing technique [79].

### Conclusions and Outlook for the Future

The preceding sections of the chapter have really only glossed over the many topics related to noise and instability in lasers, and to modelocked soliton lasers, specifically. The author felt that a deeper appreciation could be gained if the reader were exposed to some of the details of fundamental noise sources like shot noise, the Schawlow–Townes linewidth, the master modelocking equation, as well as the Lorenz–Haken equations. These form the basis for most of the serious studies of noise and instability today. While already long, this chapter has largely neglected topics having to do with fiber lasers [122,123], issues of polarization properties of vector solitons in modelocked fiber lasers and instabilities called “exploding solitons” [124], spatial instabilities [16,23] and instabilities in semiconductor lasers [23]. These omissions were made not because they lacked merit, but because space was fundamentally limited. To those scientists in their respective disciplines, I apologize.

In just the past few years, the field has seen a reinvigoration with new ideas and a quest for determining theoretically and, hopefully, demonstrating experimentally, the ultimate limits in modelocked laser stability. New results on pump power fluctuations causing relaxation oscillations and frequency pulling can be found in [67]. The real “holy grail”, of course, would be a quantum mechanically correct expression for the frequency stability or, equivalently, the modal linewidth of a modelocked soliton laser, and which would be readily accessible to experimental verification.

### Acknowledgments

The author is greatly indebted to many colleagues, friends and teachers who have influenced me over the years. I wish to thank Dr. Steven T. Cundiff and Professor Curtis R. Menyuk for their tremendous contributions and stimulating conversations regarding their recent efforts in understanding this vast field. To Dr. John L. Hall I owe a great and hearty thanks for all of his pioneering work, for his patience as a sounding board and his unique perspective on physics. Thanks also to Professors Anthony E. Siegman, Stephen E. Harris, Robert L. Byer, and especially, Alwyn C. Scott, from whom I learned much. Finally, on behalf of my group at the University of California, I wish to thank Robert Temple, Tom Faulkner and Roger Muat of Agilent Technologies (formerly Hewlett-Packard Company) for numerous donations and countless hours of tutoring. This work was supported in part by the David and Lucile Packard Foundation and the National Science Foundation under grant ECS-0622235.

### Bibliography

1. Barnes JA, Chi AR, Cutler LS, Healey DJ, Leeson DB, McGunigal TE, Mullen JA Jr, Smith WL, Sydnor RL, Vessot RFC, Winkler GMR (1971) Characterization of frequency stability. *IEEE Trans Instrum Meas* IM-20:105–120
2. Kroupa VF (ed) (1983) *Frequency stability: Fundamentals and measurement*. IEEE Press, New York
3. Sullivan DB, Allan DW, Howe DA, Walls FL (eds) (1990) *Characterization of clocks and oscillators: NIST Technical Note 1337*. United States Government Printing Office, Washington
4. Siegman AE (1986) *Lasers*. University Science Books, Mill Valley
5. Milonni PW, Shih M, Ackerhalt JR (1987) *Chaos in Laser-Matter Interactions*. World Scientific, Singapore
6. Narducci LM, Abraham NB (1988) *Laser Physics and Laser Instabilities*. World Scientific, Singapore
7. Saleh BEA, Teich MC (1991) *Fundamentals of Photonics*. Wiley, New York
8. Svanberg S (1992) *Atomic and Molecular Spectroscopy*, 2nd edn. Springer, Berlin
9. Demtröder W (1996) *Laser Spectroscopy*, 2nd edn. Springer, Berlin
10. Yariv A (1997) *Optical Electronics in Modern Communications*, 5th edn. Oxford University Press, Oxford
11. Koehnner W (1996) *Solid State Laser Engineering*, 4th edn. Springer, Berlin
12. Haken H (1985) Light. In: *Laser Dynamics*, vol 2. North-Holland, Amsterdam
13. Boyd RW, Raymer MG, Narducci LM (eds) (1986) *Optical Instabilities*. In: *Proceedings of the International Meeting on Instabilities and Dynamics of Lasers and Nonlinear Optical Systems*, University of Rochester, 18–21 June, 1985. Cambridge University Press, Cambridge
14. Abraham NB, Mandel P, Narducci LM (1988) Dynamical instabilities and pulsations in lasers. In: Wolf E (ed) *Progress in Optics*, vol XXV, ch. I. North-Holland, Amsterdam, pp 1–190



15. Robins WP (1982) *Phase Noise in Signal Sources*. Peter Peregrinus, London
16. Weiss CO, Vilaseca R (1991) *Dynamics of Lasers*. Weinheim, New York
17. Riehle F (2004) *Frequency Standards*. Wiley-VCH, Weinheim
18. Kärtner FX, Morgner U, Schibli T, Ell R, Haus HA, Fujimoto JG, Ippen EP (2004) Few-Cycle Pulses Directly from a Laser. In: *Topics in Applied Physics*, vol 95. Springer, Berlin, pp 73–136
19. Ye J, Cundiff ST (eds) (2005) *Femtosecond Optical Frequency Comb Technology: Principle, Operation and Application*. Springer, New York
20. Jackson JD (1999) *Classical Electrodynamics*, 3rd edn. Wiley, New York
21. Slichter CP (1963) *Principles of Magnetic Resonance*. Harper and Row, New York
22. Hecht E (2002) *Optics*, 4th edn. Addison-Wesley, San Francisco
23. van Tartwijk GHM, Agrawal GP (1998) Laser instabilities: a modern perspective. *Prog Quant Elect* 22:43–122
24. Dunsmuir R (1961) Theory of relaxation oscillations in optical masers. *J Electron Contr* 10:453–458
25. Begon M, Harper JL, Townsend CR (1986) *Ecology: Individuals, Populations, and Communities*. Sinauer Assoc, Sunderland
26. Davis HT (1962) *Introduction to Nonlinear Differential and Integral Equations*. Dover, New York
27. Kaplan JI, Zier R (1962) Model for transient oscillations in a three-level optical maser. *J Appl Phys* 33:2372–2375
28. Bostick HA, O'Connor JR (1962) Infrared oscillations from  $\text{CaF}_2:\text{U}^{+3}$  and  $\text{BaF}_2:\text{U}^{+3}$  masers. *Proc IRE* 50:219–220
29. Tang CL (1963) On maser rate equations and transient oscillations. *J Appl Phys* 34:2935–2940
30. Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20:130–141
31. Gleick J (1987) *Chaos*. Penguin, New York
32. Schuster HG, Just W (2005) *Deterministic Chaos*, 4th edn. Wiley-VCH, Weinheim
33. Haken H (1975) Analogy between higher instabilities in fluids and lasers. *Phys Lett* 53A:77–78
34. Casperson LW (1978) Spontaneous coherent pulsations in laser oscillators. *IEEE J Quantum Electron* QE-14:756–761
35. Mayr M, Risken H, Vollmer HD (1981) Periodic and chaotic breathing of pulses in a ringlaser. *Optics Comm* 36:480–482
36. Weiss CO, Klische W (1984) On observability of Lorenz instabilities in lasers. *Optics Comm* 51:47–48
37. Lugiato LA, Narducci LM, Bandy DK, Pennise CA (1983) Breathing, spiking and chaos in a laser with injected signal. *Opt Commun* 46:64–68
38. Weiss CO, Godone A, Olafsson A (1983) Routes to chaotic emission in a cw He-Ne laser. *Phys Rev A* 28:892–895
39. Arrechi FT, Lippi GL, Puccioni GP, Tredicce JR (1984) Deterministic chaos in laser with injected signal. *Optics Comm* 51:308–314
40. Shih M, Milonni PW, Ackerhalt JR (1985) Modeling laser instabilities and chaos. *J Opt Soc Amer B* 2:130–135
41. Narducci LM, Sadiqy H, Lugiato LA, Abraham NB (1985) Experimentally accessible periodic pulsations of a single-mode homogeneously broadened laser (the Lorenz model). *Optics Comm* 55:370
42. Weiss CO, Brock J (1986) Evidence for Lorenz-type chaos in a laser. *Phys Rev Lett* 57:2804–2806
43. Pujol J, Laguarda F, Vilaseca R, Corbalán R (1988) Influence of pump coherence on the dynamic behavior of a laser. *J Opt Soc Amer B* 5:1004–1010
44. Lauterborn W, Steinhoff R (1988) Bifurcation structure of a laser with pump modulation. *J Opt Soc Amer B* 5:1097–1104
45. Brunner W, Fischer R, Paul H (1988) Time evolution of the total electric-field strength in multimode lasers. *J Opt Soc Amer B* 5:1139–1143
46. Khanin YI (1988) Mechanisms of nonstationary behavior of solid-state lasers. *J Opt Soc Amer B* 5:889–898
47. Abraham NB, Allen UA, Peterson E, Vicens A, Vilaseca R, Espinosa V, Lippi GL (1995) Structural similarities and differences among attractors and their intensity maps in the Laser-Lorenz model. *Optics Comm* 117:367–384
48. Smith CP, Dykstra R (1996) Observation in the two-level spatial Maxwell-Bloch model of the anomalously large first peak as seen in experimental Lorenz-like spiral chaos from the  $^{15}\text{NH}_3$  laser. *Optics Comm* 129:69–74
49. Lauterborn W, Kurz T (2003) *Coherent Optics*, 2nd edn. Springer, Berlin
50. Mørk J, Mark J, Tromborg B (1990) Route to chaos and competition between relaxation oscillations for a semiconductor laser with optical feedback. *Phys Rev Lett* 65:1999–2002
51. Hentschel M, Kienberger R, Spielmann C, Reider GA, Milosevic N, Brabec T, Corkum P, Heinzmann U, Drescher M, Krausz F (2001) Attosecond metrology. *Nature* 414:509–513
52. Diels J-C, Rudolf W (1996) *Ultrashort Laser Pulse Phenomena*. Academic Press, San Diego
53. Scott A (2003) *Nonlinear Science; Emergence and Dynamics of Coherent Structures*, 2nd edn. Oxford University Press, Oxford
54. Korteweg DJ, deVries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philos Mag Ser 5* 39:422–443
55. Russell JS (1838) Report of the committee on waves. Report of the 7th Meeting of the British Association for the Advancement of Science, pp 417–496
56. Lamb GL Jr (1980) *Elements of Soliton Theory*. Wiley, New York
57. Ablowitz MJ, Segur H (1981) *Solitons and the Inverse Scattering Transform*. Society for Industrial and Applied Mathematics, Philadelphia
58. Zabusky NJ, Kruskal MD (1965) Interaction of “solitons in a collisionless plasma and the recurrence of initial states”. *Phys Rev Lett* 15:240–243
59. Scott AC, Chu FYF, McLaughlin DW (1973) The soliton – a new concept applied science. *Proc IEEE* 61:1443–1483
60. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers. I. Anomalous dispersion. *Appl Phys Lett* 23:142–144
61. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. *Phys Rev Lett* 45:1095–1098
62. Haus HA (1975) Theory of mode locking with a fast saturable absorber. *J Appl Phys* 46:3049–3058
63. Haus HA, Fujimoto JG, Ippen EP (1991) Structures for additive pulse modelocking. *J Opt Soc Amer B* 8:2068–2076
64. Haus HA, Fujimoto JG, Ippen EP (1992) Analytic theory of additive pulse and Kerr lens mode locking. *IEEE J Quantum Electron* 28:2086–2096

65. Haus HA, Mecozzi A (1993) Noise of mode-locked lasers. *IEEE J Quantum Electron* 29:983–996
66. Kapitula T, Kutz JN, Sandstede B (2002) Stability of pulses in the master mode-locking equation. *J Opt Soc Amer B* 19:740–746
67. Menyuk CR, Wahlstrand JK, Willits J, Smith RP, Schibli TR, Cundiff ST (2007) Pulse dynamics in mode-locked lasers: relaxation oscillations and frequency pulling. *Opt Express* 15:6677–6689
68. Papoulis A (1965) *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York
69. Cooper GR, McGillem CD (1971) *Probabilistic Methods of Signal and System Analysis*, 2nd edn. Holt, Rinehart and Winston, Fort Worth
70. Schawlow AL, Townes CH (1958) Infrared and optical masers. *Phys Rev* 112:1940–1949
71. Salomon C, Hils D, Hall JL (1988) Laser stabilization at the millihertz level. *J Opt Soc Amer B* 5:1576–1587
72. Eichenseer M, von Zanthier J, Walther H (2005) Common-mode-free frequency comparison of lasers with relative frequency stability at the millihertz level. *Opt Lett* 30:1662–1664
73. Notcutt M, Ma L-S, Hall JL (2005) Simple and compact 1 Hz laser system via an improved mounting configuration of a reference cavity. *Opt Lett* 30:1815–1817
74. Loudon R (1973) *The Quantum Theory of Light*. Oxford University Press, Oxford
75. Kingston RH (1995) *Optical Sources, Detectors, and Systems*. Academic Press, San Diego
76. Feller W (1957) *An Introduction to Probability Theory and Its Applications*. Wiley, New York, p 168
77. Gradshteyn IS, Ryzhik IM (1980) *Table of Integrals, Series, and Products*, corrected and enlarged 4th edn. Academic Press, New York
78. Abramowitz M, Stegun IA (eds) (1965) *Handbook of Mathematical Functions*, 55. US Department of Commerce, National Bureau of Standards, Washington
79. Scott RP, Langrock C, Kolner BH (2001) High dynamic range laser amplitude and phase noise measurement techniques. *IEEE J Select Topics Quant Electron* 7:641–655
80. Kluge J, Wiechert D, Linde DV (1984) Fluctuations in synchronously modelocked dye lasers. *Opt Commun* 51:271–277
81. Helbing FW, Steinmeyer G, Keller U, Windeler RS, Stenger J, Telle HR (2002) Carrier-envelope offset dynamics of mode-locked lasers. *Opt Lett* 27:194–196
82. Scott RP, Kolner BH, Langrock C, Byer RL, Fejer MM (2003) Ti:sapphire laser pump-noise transfer function. In: *Proceedings of the Conference on Lasers and Electro-optics*, Paper CFB2, Baltimore
83. Kolner BH, Scott RP, Langrock C (2003) Laser phase noise degradation from thermal effects due to pump power fluctuations. In: *Proceedings of the 2003 IEEE/LEOS Summer Topical Meeting on Photonic Time/Frequency Measurement and Control*, Paper TuC3.3, Vancouver, Institute of Electrical and Electronics Engineers, 14–16 July
84. Mulder TD, Scott RP, Baker KA, Kolner BH (2007) Characterization of the complex noise transfer function of a modelocked Ti:sapphire laser. In: *Conference on Lasers and Electro-Optics (CLEO 2007)*, Baltimore, Paper JThD38
85. Scott RP, Mulder TD, Baker KA, Kolner BH (2007) Amplitude and phase noise sensitivity of modelocked Ti:sapphire lasers in terms of a complex noise transfer function. *Opt Express* 15:9090–9095
86. Weber R, Neuenschwander B, Donald MM, Roos MB, Weber HP (1998) Cooling schemes for longitudinally diode laser-pumped Nd:YAG rods. *IEEE J Quantum Electron* 34:1046–1053
87. Kolner BH (2008) Dynamic temperature distribution in cylindrical laser rods with time-varying pump sources. (in preparation)
88. Kolner BH, Mulder TD, Scott RP (2008) Laser noise modulation transfer functions. (in preparation)
89. Mulder TD, Scott RP, Kolner BH (2008) Amplitude and envelope phase noise of a modelocked laser predicted from its noise transfer function and the pump noise power spectrum. *Opt Express* 16(18):14186–14191
90. Bender CM, Orszag SA (1978) *Advanced Mathematical Methods for Scientists and Engineers*. McGraw-Hill, New York
91. Haus HA, Lai Y (1990) Quantum theory of soliton squeezing: a linearized approach. *J Opt Soc Amer B* 7:386–392
92. Haberman R (1987) *Elementary Applied Partial Differential Equations with Fourier Series and Boundary Value Problems*, 2nd edn. Prentice-Hall, New Jersey
93. Kuizenga DJ, Siegman AE (1970) FM and AM mode locking of the homogeneous laser – Part I: Theory. *IEEE J Quantum Electron* QE-6:694–708
94. Kuizenga DJ, Siegman AE (1970) FM and AM mode locking of the homogeneous laser – Part II: Experimental results in a Nd:YAG laser with internal FM modulation. *IEEE J Quantum Electron* QE-6:709–715
95. Scavennec A (1976) Mismatch effects in synchronous pumping of the continuously operated mode-locked laser. *Optics Comm* 17:14–17
96. Zheng JP, Sen U, Benenson DM, Kwok HS (1986) Observation of periodicity multiplication in a synchronously pumped dye laser. *Opt Lett* 11:632–634
97. MacFarlane DL, Casperson LW (1987) Pulse-train instabilities in a mode-locked argon laser: Experimental studies. *J Opt Soc Amer B* 4:1777–1780
98. MacFarlane DL, Casperson LW, Tovar AA (1988) Spectral behavior and pulse train instabilities of a synchronously pumped mode-locked dye laser. *J Opt Soc Amer B* 5:1144–1152
99. Scott RP, Bennett CV, Kolner BH (1997) AM and high-harmonic FM modelocking. *Appl Opt* 36:5908–5912
100. Valdmantis JA, Fork RL, Gordon JP (1985) Generation of optical pulses as short as 27 femtoseconds directly from a laser balancing self-phase modulation, group-velocity dispersion, saturable absorption, and saturable gain. *Opt Lett* 10:131–133
101. Avramopoulos H, French PMW, Williams JAR, New GHC, Taylor JR (1988) Experimental and theoretical studies of complex pulse evolutions in a passively mode-locked ring dye laser. *IEEE J Quantum Electron* 24:1884–1892
102. Spence DE, Kean PN, Sibbett W (1991) 60-fsec pulse generation from a self-modelocked Ti:sapphire laser. *Opt Lett* 16:42–44
103. Christov IP, Kapteyn HC, Murnane MM, Huang C-P, Zhou J (1995) Space-time focusing of femtosecond pulses in a Ti:sapphire laser. *Opt Lett* 20:309–311
104. Kalosha VP, Müller M, Herrmann J, Gatz S (1998) Spatiotemporal model of femtosecond pulse generation in Kerr-lens mode-locked solid-state lasers. *J Opt Soc Amer B* 15:535–550

105. Christov IP, Stoev VD (1998) Kerr-lens mode-locked laser model: role of space time effects. *J Opt Soc Amer B* 15:1960–1966
106. Jirauschek C, Kärtner FX, Morgner U (2003) Spatiotemporal Gaussian pulse dynamics in Kerr-lens mode-locked lasers. *J Opt Soc Amer B* 20:1356–1368
107. Sergeev AM, Vanin EV, Wise FW (1997) Stability of passively modelocked lasers with fast saturable absorbers. *Optics Comm* 140:61–64
108. Liu Y-M, Prucnal PR (1993) Slow amplitude modulation in the pulse train of a self-modelocked Ti:sapphire laser. *IEEE J Quantum Electron* 29:2663–2669
109. Sucha G, Bolton SR, Weiss S, Chemla DS (1995) Period doubling and quasi-periodicity in additive-pulse mode-locked lasers. *Opt Lett* 20:1794–1796
110. Sánchez LM, Hnilo AA (2001) Description of Kerr lens mode-locked lasers with Poincaré maps in the complex plane. *Optics Comm* 199:189–199
111. Hall JL, Ye J, Diddams SA, Ma L-S, Cundiff ST, Jones DJ (2001) Ultrasensitive spectroscopy, the ultrastable lasers, the ultrafast lasers, and the seriously nonlinear fiber: a new alliance for physics and metrology. *IEEE J Quantum Electron* 37:1482–1492
112. Holzwarth R, Zimmermann M, Udem T, Hänsch TW (2001) Optical clockworks and the measurement of laser frequencies with a mode-locked frequency comb. *IEEE J Quantum Electron* 37:1493–1501
113. <http://nobelprize.org/physics/laureates/2005/>
114. Cundiff ST, Ye J (2003) Femtosecond optical frequency combs. *Rev Modern Phys* 75:325–342
115. Helbing FW, Steinmeyer G, Stenger J, Telle HR, Keller U (2002) Carrier-envelope-offset dynamics and stabilization of femtosecond pulses. *Appl Phys B* 74:S34–S42
116. Ranka J, Windeler R, Stentz A (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Opt Lett* 25:25–27
117. Ranka JK, Windeler RS, Stentz AJ (2000) Optical properties of high-delta air silica microstructure optical fibers. *Opt Lett* 25:796–798
118. Xu L, Spielmann C, Poppe A, Brabec T, Krausz F, Hänsch TW (1996) Route to phase control of ultrashort light pulses. *Opt Lett* 21:2008–2010
119. Matos L, Mücke OD, Jian C, Kärtner FX (2006) Carrier-envelope phase dynamics and noise analysis in octave-spanning Ti:sapphire lasers. *Opt Express* 14:2497–2511
120. Stenger J, Talle HR (2000) Intensity-induced mode shift in a femtosecond laser by a change in the nonlinear index of refraction. *Opt Lett* 25:1553–1555
121. Holman KW, Jones RJ, Marian A, Cundiff ST, Ye J (2003) Detailed studies and control of intensity related dynamics of femtosecond frequency combs from mode-locked Ti:sapphire lasers. *IEEE J Select Topics Quantum Electron* 9:1018
122. Kutz JN (1998) Modelocking pulse dynamics in fiber lasers. In: *SPIE Conference on Physics and Simulation of Optoelectronics Devices VI*, vol 3283, (San Jose, CA). SPIE, pp 639–651
123. Washburn BR, Swan WC, Newberry NR (2005) Response dynamics of the frequency comb output from a femtosecond fiber laser. *Opt Lett* 13:10622–10633
124. Cundiff ST (2005) Soliton dynamics in mode-locked lasers. *Lecture notes in physics*, vol 661. Springer, Berlin, pp 183–206

## Non-linear Dynamics, Symmetry and Perturbation Theory in

GIUSEPPE GAETA

Dipartimento di Matematica, Università di Milano,  
Milan, Italy

### Article Outline

Glossary

Definition of the Subject

Introduction

Symmetry of Dynamical Systems

Perturbation Theory: Normal Forms

Perturbative Determination of Symmetries

Symmetry Characterization of Normal Forms

Symmetries and Transformation to Normal Form

Generalizations

Symmetry for Systems in Normal Form

Linearization of a Dynamical System

Further Normalization and Symmetry

Symmetry Reduction of Symmetric Normal Forms

Conclusions

Future Developments

Additional Notes

Bibliography

### Glossary

**Perturbation theory** A theory aiming at studying solutions of a differential equation (or system thereof), possibly depending on external parameters, near a known solution and/or for values of external parameters near to those for which solutions are known.

**Dynamical system** A system of first order differential equations  $dx^i/dt = f^i(x, t)$ , where  $x \in M$ ,  $t \in \mathbf{R}$ . The space  $M$  is the phase space for the dynamical system, and  $\tilde{M} = M \times \mathbf{R}$  is the extended phase space. When  $f$  is smooth we say the dynamical system is smooth, and for  $f$  independent of  $t$ , we speak of an autonomous dynamical system.

**Symmetry** An invertible transformation of  $\tilde{M}$  mapping solutions into solutions. If the dynamical system is smooth, smoothness will also be required on symmetry transformations; if it is autonomous, it will be natural to consider transformations of  $M$  rather than of  $\tilde{M}$ .

**Symmetry reduction** A method to reduce the equations under study to simpler ones (e.g. with less dependent variables, or of lower degree) by exploiting their symmetry properties.

**Normal form** A convenient form to which the system of differential equations under study can be brought by

means of a sequence of change of coordinates. The latter are in general well defined only in a subset of  $M$ , possibly near a known solution for the differential equations.

**Further normalization** A procedure to further simplify the normal form for a dynamical system, in general making use of certain degeneracies in the equations to be solved in the course of the normalization procedure.

### Definition of the Subject

Given a differential equation or system of differential equations  $\Delta$  with independent variables  $\xi^a \in \mathcal{E} \subseteq \mathbb{R}^q$  and dependent variables  $x^i \in M \subseteq \mathbb{R}^p$ , a symmetry of  $\Delta$  is an invertible transformation of the extended phase space  $\widetilde{M} = \mathcal{E} \times M$  into itself which maps solutions of  $\Delta$  into (generally, different) solutions of  $\Delta$ .

The presence of symmetries is a non-generic feature; correspondingly, equations with symmetry have some special features. These can be used to obtain information about the equation and its solutions, and sometimes allow one to obtain explicit solutions.

The same applies when we consider a perturbative approach to the equations: taking into account the presence of symmetries guarantees the perturbative expansion has certain specific features (e. g. some terms are not allowed) and hence allows one to deal with simplified expansions and equations; thus this approach can be of great help in providing explicit solutions.

As mentioned above, symmetry is a non-generic feature: if we take a “generic” equation or system, it will not have any symmetry property. What makes the symmetry approach useful and widely applicable is a remarkable fact: many of the equations encountered in applications, and especially in physical and related ones (mechanical, electronic, etc.) are symmetric; this in turn descends from the fact that the fundamental equations of physics have a high degree of symmetry.

Thus, symmetry-based methods are at the same time “non-generic” in a mathematical sense, and “general” in a physical, or more generally real-world, sense.

### Introduction

Symmetry has been a major ingredient in the development of quantum perturbation theory, and is a fundamental ingredient of the theory of integrable (Hamiltonian and non-Hamiltonian) systems; yet, the use of symmetry in the context of general perturbation theory is rather recent.

From the point of view of nonlinear dynamics, the use of symmetry has become widespread only through equiv-

ariant bifurcation theory; even in this case, attention has been mostly confined to linear symmetries.

Also, in recent years the theory and practice of symmetry methods for differential equations became increasingly popular and has been applied to a variety of problems (to a large extent, following the appearance of the book by Olver [151]). This theory is deeply geometrical and deals with symmetries of general nature (provided that they are described by smooth vector fields), i. e. in this context there is no reason to limit attention to linear symmetries.

In this article we look at the basic tools of perturbation theory, i. e. *normal forms* (first introduced by Poincaré more than a century ago for general dynamical systems; the Hamiltonian case being studied in its special features by Birkhoff several decades ago) and study their interaction with symmetries, with no limitation to linear ones. See the articles ► [Normal Forms in Perturbation Theory](#), ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#) for an introduction to Normal Forms.

We focus on the most basic setting, i. e. systems having a fixed point (at the origin) and perturbative expansions around this; thus our theory is entirely local. We also limit to the discussion of general vector fields, i. e. we will not discuss the formulation one would obtain for the special case of Hamiltonian vector fields ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#), [111] (in which case one can deal with the Hamiltonian function rather than with the vector field it generates), referring the reader to [51] for this as well as for other extensions and for several proofs.

We start by recalling basic notions about the symmetry of differential equations, and in particular of dynamical systems; we will then discuss normal forms in the presence of symmetries, and the problem of taking into normal form the dynamical vector field and the symmetry vector field(s) at the same time.

The presence of symmetry causes several peculiar phenomena in the dynamics, and hence also in perturbative expansions. This has been explained in very effective terms by Ian Stewart [175]:

*Symmetries abound in nature, in technology, and – especially – in the simplified mathematical models we study so assiduously. Symmetries complicate things and simplify them. They complicate them by introducing exceptional types of behavior, increasing the number of variables involved, and making vanish things that usually do not vanish. They simplify them by introducing exceptional types of behavior, increasing the number of variables involved, and making*



vanish things that usually do not vanish. They violate all the hypotheses of our favorite theorems, yet lead to natural generalizations of those theorems. It is now standard to study the “generic” behavior of dynamical systems. Symmetry is not generic. The answer is to work within the world of symmetric systems and to examine a suitably restricted idea of genericity.

Here we deal with dynamical systems, and more specially autonomous ones, i. e. systems of equations of the form  $dx^i/dt = f^i(x)$ . Now we have a single independent variable, the time  $t \in \mathbf{R}$ , and in view of its distinguished role we will mainly focus attention on transformations leaving it unchanged.

It is appropriate to point out here connections to several topics which we will *not* illustrate in this article.

First of all, we stress that we will work at the formal level, i. e. without considering the problem of *convergence* of the power series entering in the theory. This convergence is studied in the articles ► [Perturbation Theory](#), ► [Perturbative Expansions](#), [Convergence of](#), to which the interested reader is referred in the first instance.

As hinted above, perturbation theory for symmetric systems has many points of contact with the topic of *Equivariant Bifurcation Theory*, which we will not touch upon here. The interested reader is referred to [\[104,107,118,160,179\]](#) for Bifurcation Theory in general, and then for the equivariant setting to the books [\[42,105,118\]](#). More compact introductions are provided by the review papers [\[55,79\]](#).

Many facets of the interplay of symmetry and perturbation theory are also discussed in the SPT conference proceedings volumes [\[1,14,17,57,96,98\]](#).

Our discussion is based on the treatment in [\[51\]](#), with integrations and updates where appropriate.

Some considerations and remarks are given in additional notes collected in the last section; these are called for by marks<sup>(xx)</sup> with xx consecutive numbers.

## Symmetry of Dynamical Systems

Symmetry of differential equations – and its use to solve or reduce the differential equations themselves – is a classical and venerable subject, being the very motivation to Sophus Lie when he created what is nowadays known as the theory of Lie groups [\[121\]](#). The subject is now dealt with in a number of textbooks (see e. g. [\[3,19,26,27,37,80,115,125,151,152,174\]](#)) and review papers (see e. g. [\[116,184,185,191,192\]](#)); we will thus refer the reader to these for the general theory, and briefly recall here the special formulation one obtains when dealing with symmetries of smooth dynamical systems in  $\mathbf{R}^n$ .

Consider a (possibly non-autonomous) system

$$\dot{x}^i = f^i(x; t) \quad i = 1, \dots, n; \quad (1)$$

we assume  $x \in M = \mathbf{R}^n$ ;  $M$  is also called the *phase space*, and  $\tilde{M} = M \times \mathbf{R}$  (the second factor representing of course time  $t$ ) is the *extended phase space*.<sup>(1)</sup>

We consider now vector fields in  $\tilde{M}$ ; these can be written in coordinates as

$$S = \tau(x, t) \frac{\partial}{\partial t} + \sum_{i=1}^n \varphi^i(x, t) \frac{\partial}{\partial x^i}. \quad (2)$$

Note that (1) is identified with the vector field

$$X_f := \sum_{i=1}^n f^i(x, t) \frac{\partial}{\partial x^i}. \quad (3)$$

A (vector) function  $x: \mathbf{R} \rightarrow M$  is naturally identified with the subset  $\sigma_x$  of  $\tilde{M}$  (corresponding to its graph) defined by

$$\sigma_x = \{(y, t) \in M \times \mathbf{R} : y^i = x^i(t)\} \subset \tilde{M}. \quad (4)$$

The vector field  $S$  acts infinitesimally in  $\tilde{M}$  by mapping points  $(y, t)$  to points  $(\hat{y}, \hat{t})$  given by

$$\hat{t} = t + \varepsilon \tau(y, t), \quad \hat{y}^i = y^i + \varepsilon \varphi^i(y, t); \quad (5)$$

as  $\varepsilon$  is small these relations can be inverted, yielding at first order in  $\varepsilon$

$$t = \hat{t} - \varepsilon \tau(\hat{y}, \hat{t}), \quad y^i = \hat{y}^i - \varepsilon \varphi^i(\hat{y}, \hat{t}). \quad (6)$$

Using these relations, it is easy to check that the subset  $\sigma = \sigma_x$  is mapped by  $S$  to a (generally) different subset  $\hat{\sigma}$ , corresponding to  $y = \hat{x}(t)$ , with

$$\hat{x}^i(t) = x^i(t) + \varepsilon \left[ \varphi^i(x(t), t) - \dot{x}^i(t) \tau(x(t), t) \right]. \quad (7)$$

We say that  $S$  is a **symmetry** for the dynamical system (1) if it maps solutions into (generally, different) solutions. The condition for this to happen turns out to be [\[51\]](#)

$$\frac{\partial \varphi^i}{\partial t} - \frac{\partial \tau}{\partial t} f^i = \varphi^j \frac{\partial f^i}{\partial x^j} - f^j \frac{\partial \varphi^i}{\partial x^j}. \quad (8)$$

This can be more compactly expressed by introducing the Lie–Poisson bracket

$$\{f, g\} := (f \cdot \nabla) g - (g \cdot \nabla) f \quad (9)$$

between vector functions on  $\tilde{M}$ . Then (8) reads

$$(\partial \varphi / \partial t) - (\partial \tau / \partial t) f = \{\varphi, f\}. \quad (10)$$



In the following we will consider *autonomous* dynamical systems; in this case it is rather natural to consider only transformations which leave  $t$  invariant, i. e. with  $\tau = 0$ . In this case (10) reduces to

$$(\partial\varphi/\partial t) + \{f, \varphi\} = 0. \quad (11)$$

A further reduction is obtained if we only consider transformations for which the action on  $M$  is also independent of time, so that  $\partial\varphi/\partial t = 0$  and the symmetry condition is

$$\{f, \varphi\} = 0; \quad (12)$$

in this case one speaks of *Lie-point time-independent (LPTI) symmetries*.

The Eqs. (8) (or its reductions) will be referred to as the *determining equations* for the symmetries of the dynamical system (1).<sup>(2)</sup>

It should be stressed that (8) are linear in  $\varphi$  and  $\tau$ ; it is thus obvious that the solutions will span a linear space. It is also easy to check (the proof of this fact follows from the bilinearity of (9) and the Jacobi identity) that if  $S_1$  and  $S_2$  are solutions to (8), so is their Lie–Poisson bracket  $\{S_1, S_2\}$ . The set  $\mathcal{G}_{X_f}$  of vector fields  $X_\varphi$  with  $\varphi$  solutions to (8) is thus a Lie algebra; it is the *symmetry algebra* for the dynamical system (1).

The symmetry algebra of a dynamical system is infinite dimensional, but has moreover an additional structure. That is, it is a *module* over the algebra  $\mathcal{I}_{X_f}$  of first integrals for  $f$  (that is, scalar functions  $\alpha: M \rightarrow \mathbb{R}$  such that  $X_f(\alpha) \equiv (f \cdot \nabla)\alpha = 0$ ). Albeit  $\mathcal{G}_{X_f}$  is infinite dimensional as a Lie algebra, it is not so as a Lie module. We have, indeed [186]:

**Theorem 1 (Walcher)** *The set  $\mathcal{G}_{X_f}$  is a finitely generated module over  $\mathcal{I}_{X_f}$ .*

### Perturbation Theory: Normal Forms

In this section we recall some basic facts about perturbation theory for general dynamical systems, referring to ► [Normal Forms in Perturbation Theory](#), ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#), ► [Perturbation Theory](#) for details. For the sake of simplicity, we discuss perturbations around an equilibrium point; see e. g. [6,7,8,111,160,164,165] for more general settings.

As is well known – and discussed, e. g. in ► [Normal Forms in Perturbation Theory](#), ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#) – a central objective of perturbation theory is to set (1) in normal form, i. e. to eliminate as many nonlinear terms as possible, so that the difference with respect to the linearized equation is as

small as possible (see again ► [Normal Forms in Perturbation Theory](#), ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#), or ► [Perturbation Theory](#), for a precise meaning to this statement; a lengthier discussion is given e. g. in [5,7,99,104,179]).<sup>(3)</sup>

We briefly recall how this goes, also in order to fix notation. We consider a  $C^\infty$  dynamical system  $\dot{x} = f(x)$  in  $\mathbb{R}^n$ , admitting  $x = 0$  as an equilibrium point – that is with  $f(0) = 0$ . By Taylor-expanding  $f(x)$  around  $x = 0$  we will write this in the form

$$\dot{x} = f(x) = \sum_{k=0}^{\infty} F_k(x) \quad (13)$$

where  $F_k(x)$  is homogeneous of degree  $(k+1)$  in  $x$  (this seemingly odd notation will come out handy in the following). We denote the linear space of vector function  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$  homogeneous of degree  $(k+1)$  by  $\mathcal{V}_k$ .

### Poincaré–Dulac Normal Forms

Let us consider a change of coordinates of the form

$$\xi^i = x^i + h_k^i(x), \quad (14)$$

where  $h_k \in \mathcal{V}_k (k \geq 1)$ ; we write  $\Gamma_j^i = (\partial h^i / \partial x^j)$ . A change of coordinates of the form (14) is called a *Poincaré transformation* (or P transformation for short), and the function  $h_k$  is also called the *generator* of the P transformation. In the following we will freely drop the subscript  $k$  when this cannot generate any confusion.

The transformation (14) is, for small  $x$ , a near-identity transformation; thus it is surely invertible in a small enough neighborhood of the origin. We apply  $\Lambda := (I + \Gamma)^{-1}$  on (13), and get the P transformed dynamical system in the form

$$\dot{x} = \tilde{f}(x) \equiv \Lambda \sum_{m=0}^{\infty} F_m(x + h_k(x)) = \sum_{m=0}^{\infty} \tilde{F}_m(x). \quad (15)$$

In order to identify the  $\tilde{F}_m$  we should consider power series expansions for  $\Lambda$  and for  $F_m(x + h(x))$ . With standard computations (we refer again to ► [Normal Forms in Perturbation Theory](#), or to [7,51], for details), we obtain that the  $\tilde{F}_m$  are given (with  $[q]$  the integer part of  $q$ ) by

$$\tilde{F}_m = F_m + \sum_{p=1}^{[m/k]} \left[ \sum_{s=0}^p (-1)^s \Gamma^s \Phi_{h_k}^{p-s} \right] F_{m-kp}. \quad (16)$$

The  $\Phi_h^r$  appearing in (16) are defined as follows. With a multi-index notation, write  $J = (j_1, \dots, j_n)$ ,  $|J| = \sum_i j_i$ ; set then  $\partial_J := \partial_1^{j_1} \dots \partial_n^{j_n}$ , and similarly  $h_k^J :=$

$(h_k^1)^{j_1} \dots (h_k^n)^{j_n}$ . The operators  $\Phi_h^r$  (representing all the partial derivatives of order  $|J|$ ) are defined as  $\Phi_h^r = (1/r!) \sum_{|J|=r} (h^J \cdot \partial_J)$ .

Some special cases following from this general formula should be noted. As well known, the terms of degree smaller than  $k$  are not changed at all, i.e.  $\widetilde{F}_m = F_m$  for  $m < k$ , and the term of degree  $k$  is changed (writing  $h \equiv h_k$ ) according to

$$\widetilde{F}_k = F_k + [\Phi_h - \Gamma] F_0. \quad (17)$$

(Similarly, for  $0 < \nu < k$ , the term of degree  $k + \nu$  is changed into  $\widetilde{F}_{k+\nu} = F_{k+\nu} + [\Phi_h - \Gamma] F_\nu$ .) Define now, recalling (9), the operators

$$L_k = \{F_k, \cdot\}; \quad (18)$$

note  $L_k: \mathcal{V}_m \rightarrow \mathcal{V}_{m+k}$ .

The operator  $\mathcal{A} = L_0$ , associated with the linear part  $A = (Df)(0)$  of  $f$  (that is,  $F_0(x) = Ax^{(4)}$ ) is called the *homological operator*; it leaves the spaces  $\mathcal{V}_k$  invariant, and hence it admits the decomposition  $\mathcal{A} = \bigoplus_{m=0}^{\infty} \mathcal{A}^{(m)}$ , where  $\mathcal{A}^{(m)}$  is just the restriction of  $\mathcal{A}$  to  $\mathcal{V}_m$ .

In the following, we will need to consider the adjoint  $\mathcal{A}^+$  of the operator  $\mathcal{A}$ . For this we need to introduce a scalar product in the space  $\mathcal{V} = \bigcup \mathcal{V}_k$ . Actually, we can introduce a scalar product in each of the spaces  $\mathcal{V}_k$  into which  $\mathcal{V}$  decomposes.

A convenient scalar product was introduced in [60] (following [18]); we will only use this one (the reader should be warned that different definitions are also considered in the literature [5,51]). We denote by  $\xi_{\mu,i}$  the vector function whose components are all zero but the  $i$ th one, given by  $x^\mu := x_1^{\mu_1} \dots x_n^{\mu_n}$ ; with this notation, we define<sup>(5)</sup>

$$(\xi_{\mu,i}, \xi_{\nu,j}) = \sum_{i=1}^n \langle \xi_{\mu,i}, \xi_{\nu,i} \rangle, \quad (19)$$

where  $\langle x^\mu, x^\nu \rangle \equiv \partial^\mu x^\nu$ , and  $\partial^\mu = \partial_1^{\mu_1} \dots \partial_n^{\mu_n}$ . (When in the following we consider adjoint operators, these will be understood in terms of this.)

With this scalar product, one has the following lemma (a proof is given, e.g. in [118]).

**Lemma 1** *If  $\mathcal{A}$  is the homological operator associated with the matrix  $A$ ,  $\mathcal{A} = \{Ax, \cdot\}$ , then its adjoint  $\mathcal{A}^+$  is the homological operator associated with the adjoint matrix  $A^+$ , i.e.,  $\mathcal{A}^+ = \{A^+x, \cdot\}$ .*

We will also consider the projection onto the range of  $\mathcal{A}^{(k)}$ , denoted by  $\pi_k$ . The general homological equations

(the one for  $k = 0$  corresponds to the standard homological equation) are then

$$\mathcal{A}^{(k)}(h_k) = \pi_k F_k. \quad (20)$$

These are equations for  $h_k \in \mathcal{V}_k$ , and always admit a solution (thanks to the presence of the projection operator  $\pi_k$ ). The Eq. (20) maps into a set of algebraic equations once we introduce a basis in  $\mathcal{V}_k$ .

The  $h_k \in \mathcal{V}_k$  solving to (20) will be of the form  $h_k = h_k^* + \ell_k$ , where  $h_k^* = \mathcal{A}^*(\pi_k F_k) \in \text{Ran}[(\mathcal{A}^{(k)})^*]$  (here  $\mathcal{A}^*$  is the pseudo-inverse to  $\mathcal{A}$ ; note  $\text{Ker}(\mathcal{A}^*) = \text{Ker}(\mathcal{A}^+)$ ) is unique, and  $\ell_k$  is any function in  $\text{Ker}[\mathcal{A}^{(k)}]$ .

*Remark 1* It should be stressed that, while adding a nonzero  $\ell_k$  to  $h_k^*$  does not change the resulting  $\widetilde{F}_k$ , it could – and in general will – affect the terms of higher order.

One can then normalize  $X_f$  in the standard recursive way, based on solving homological equations; this is described, e.g. in [► Normal Forms in Perturbation Theory](#). In this way, we are reduced to considering only systems with

$$F_k \in \left[ \text{Ran} \left( \mathcal{A}^{(k)} \right) \right]^\perp = \text{Ker} \left[ \left( \mathcal{A}^{(k)} \right)^+ \right]. \quad (21)$$

Such terms are also called *resonant*.<sup>(6)</sup>

The presence of resonant terms is related to the existence of *resonance relations* among eigenvalues  $\lambda_i$  of the matrix  $A$  describing the linear part of the system; these are relations of the form  $(m \cdot \lambda) \equiv \sum_i m_i \lambda_i = \lambda_s$  where the  $m_i$  are non-negative integers, with  $|m| = \sum m_i > 1$  (the restriction  $|m| > 1$  is to avoid trivial cases); the integer  $|m|$  is also called the *order* of the resonance ([► Normal Forms in Perturbation Theory](#), [7,8]).

If the system  $\dot{x} = f(x)$  has only resonant nonlinear terms, we say that it is in *Poincaré–Dulac normal form* ([► Normal Forms in Perturbation Theory](#), [7,8,35,44]). If all the nonlinear terms of order up to  $q$  are resonant, we say that the system is in normal form up to order  $q$ .

**Theorem 2 (Poincaré–Dulac)** *Any analytic vector field  $X_f$  with  $f(0) = 0$  can be formally taken to normal form by a sequence of Poincaré transformations.*

*Remark 2* If we do not have an exact resonance, but  $(m \cdot \lambda) - \lambda_s \simeq 0$ , we have a *small denominator*, and correspondingly a very large coefficient in  $h_k$ , where  $k = |m|$ . Such small denominators are responsible for divergencies in the normalizing series [7,8,9,32,33,35,172].

## Lie Transforms

In discussing Poincaré normal forms, we have considered near-identity diffeomorphisms of  $M$ ; these can be expressed as time-one maps for the flow under some vector fields  $X_h$ . In a number of cases, it is more convenient to deal directly with such vector fields. We are going to briefly discuss this approach, and its relation with the one discussed above; for further detail, see e. g. (► [Normal Forms in Perturbation Theory](#), [25,28,29,30,51,61,84,101,145]).

Let the vector field  $H \equiv X_h$  be given, in the  $x$  coordinates, by  $H = h^i(x)(\partial/\partial x^i)$ . We denote by  $\Psi(s; x)$  the local flow under  $H$  starting at  $x$ , so that  $(d/ds)\Psi(s; x) = H(\Psi(s; x))$ . We also use exponential notation:  $\Psi(s; x) = e^{sH}x$ .

We will denote  $\Psi(1; \xi)$  as  $x$ ; the direct and inverse changes of coordinates will be defined as

$$\begin{aligned} x &= \Psi(1; \xi) = [e^{sH}\xi]_{s=1}, \\ \xi &= \Psi(-1; x) = [e^{-sH}x]_{s=1}. \end{aligned} \quad (22)$$

Now, consider another vector field  $X$  on  $M$ , describing the dynamical system we are interested in. If we study the dynamical system  $\dot{\xi}^i = f^i(\xi)$ , we consider the vector field given in the  $x$  coordinates by

$$X = X_f = f^i(x)(\partial/\partial x^i). \quad (23)$$

This also generates a (local) flow, i. e. for any  $\xi_0 \in M$  we have a one-parameter family  $\xi(t) \equiv \Gamma(t; \xi_0) \in M$  such that  $(d\xi(t)/dt) = X(\xi)$ . By means of (22), this also defines a one-parameter family  $x(t) \in M$ , which will satisfy  $(dx(t)/dt) = \tilde{X}(x)$  for some vector field  $\tilde{X}$  on  $M$ ; this will be the transformed vector field under (22), and is given by<sup>(7)</sup>

$$\tilde{X} = [e^{sH}Xe^{-sH}]_{(s=1)}. \quad (24)$$

We call this transformation the **Lie-Poincaré transformation** generated by  $h$ . Notice that this yields, up to order one in  $s$ , (and therefore if  $h \in \mathcal{V}_k$ , up to terms in  $\mathcal{V}_k$ ), just the same result as the Poincaré transformation with the same generator  $h$ .

The  $\tilde{X}$  can be given (for arbitrary  $s$ ), in terms of the Baker–Campbell–Hausdorff formula, as

$$\tilde{X} = \sum_{k=0}^{\infty} \frac{(-1)^k s^k}{k!} X^{(k)}, \quad (25)$$

where  $X^{(0)} = X$ , and  $X^{(k+1)}$  are defined recursively by  $X^{(k+1)} = [X^{(k)}, H]$ .

## Perturbative Determination of Symmetries

Let us now consider the problem of determining the symmetries of a given dynamical system. Writing the latter in the form

$$\dot{x} = f(x) = \sum_{k=0}^{\infty} F_k(x) \quad (F_k \in \mathcal{V}_k), \quad (26)$$

it is quite natural to also look for symmetries in terms of a (possibly only formal) power series; this will be our approach here.

Consider a generic smooth vector field (by smooth we mean, here and elsewhere,  $C^\infty$ ; however, we will soon go on to actually consider vector fields represented by power series) in  $M \times \mathbf{R}$ ,

$$Y = \tau(x, t) \partial_t + \sum_{i=1}^n \varphi^i(x, t) \partial_i. \quad (27)$$

Here and below,  $\partial_i \equiv (\partial/\partial x^i)$ . As remarked in Sect. “[Symmetry of Dynamical Systems](#)”, (27) is too general for our purposes; we are primarily interested in vector fields acting on  $M$  alone, and mainly [48,49,51,80] in time-independent vector fields on  $M$  (note in this way the dynamical and the symmetry vector fields are on the same footing). Thus we will just consider (we will consistently use  $X$  for the vector field  $X_f$  defined by (26), and  $Y$  for the symmetry vector field  $X_s$ , in order to simplify the notation)

$$Y = \sum_{i=1}^n s^i(x) \partial_i. \quad (28)$$

Note that at this stage we are **not** assuming the dynamical system described by  $X$  has been taken into normal form; we will see later on the specific features of this case.

## Determining Equations

Now, let us consider the dynamical system identified by  $X$ , and look for the determining equation identifying its symmetries  $Y$  as in (28). Condition (8) yields

$$f^j(x) \frac{\partial s^i(x)}{\partial x^j} - s^j(x) \frac{\partial f^i(x)}{\partial x^j} \equiv (f \cdot \nabla)s - (s \cdot \nabla)f = 0. \quad (29)$$

As discussed above, using (9) this is also written as  $\{f, s\} = 0$ , which just means (as had to be expected)

$$[X, Y] = 0. \quad (30)$$

We now denote the set of  $Y$  satisfying (30) by  $\mathcal{G}_X$ . It is obvious that  $\mathcal{G}_X$ , equipped with the usual commutator of

vector fields, is a Lie algebra. It is also easy to see that, as  $f(0) = 0$ ,  $Y \in \mathcal{G}_X$  implies  $s(0) = 0$ .

We can now expand  $Y$ , i. e.  $s(x)$ , in a perturbative series around  $x_0 = 0$  in the same way as we did for  $X$ . We write

$$s(x) = \sum_{k=0}^{\infty} S_k(x) \quad (S_k \in \mathcal{V}_k). \quad (31)$$

Plugging this into the determining Eqs. (30) we get, after rearranging the terms,

$$\sum_{k=0}^{\infty} \sum_{m=0}^k \{F_m, S_{k-m}\} = 0. \quad (32)$$

For this to hold, the different homogeneous terms of degree  $k$  must vanish separately. Thus, we have a hierarchy (in a sense to be explained in a moment) of equations

$$\sum_{m=0}^k \{F_m, S_{k-m}\} = 0 \quad k = 0, 1, 2, 3, \dots \quad (33)$$

It is convenient to isolate the terms containing linear factors, i. e. to rewrite (33) – for  $k \geq 1$  – in the form

$$\{F_0, S_k\} - \{S_0, F_k\} = - \sum_{m=1}^{k-1} \{F_m, S_{k-m}\} \equiv \chi_k, \quad (34)$$

where we have used the antisymmetry of  $\{.,.\}$ , and  $\chi_0 = \chi_1 = 0$ .

### Recursive Solution of the Determining Equations

Let us now consider the problem of concretely solving the determining Eq. (32). As the perturbative series expansion suggests, we can proceed order by order, i. e. start with consideration of the equation for  $k = 0$ , then tackle  $k = 1$ , and so on.

Proceeding in this way we are always reduced to consider equations of the form

$$F_0^j \left( \frac{\partial S_k^i}{\partial x^j} \right) = S_k^j \left( \frac{\partial F_0^i}{\partial x^j} \right) + \Psi_k^j(x) \quad (35)$$

with the  $\Psi_k^j$  known functions of  $x$  (as they depend on the known  $F_k$  and on the  $S_j$  with  $j < k$ , determined at previous stages).

Notice also that  $F_0$  is just the (known) linear part of  $X$ . If we write it in matrix form as  $F_0^i(x) = A_{ij}x^j$  (we will write similarly  $S_0^i(x) = B_{ij}x^j$  for  $S_0$ ), then (35) reads simply

$$\left[ A_{ij}x^j \right] \left( \frac{\partial S_k^i}{\partial x^j} \right) = A_{ij}S_k^j + \Psi_k^j(x). \quad (36)$$

Solving the determining equations in such a recursive way only requires one to solve at each stage a system of (inhomogeneous) linear PDEs for the  $S_k$ .

For further reference, we introduce the notation  $X_M$  for the linear vector field associated with the matrix  $M$ , i. e.  $X_M = (M_{ij}x^j)(\partial/\partial x^i)$ . We also write  $L_M$  for the homological operator associated with a matrix  $M$ ,  $L_M(.) = \{Mx, .\}$  (see also note 4). In this notation, for any two matrices  $A, B$  we have  $[X_A, X_B] = -X_{[A,B]}$ ; similarly we have, as a consequence of the Jacobi identity:

**Lemma 2** For any matrices  $A, B$ , the commutator of the associated homological operators is given by  $[L_A, L_B] = L_{[B,A]} = -L_{[A,B]}$ .

Let us follow explicitly the iterative procedure for solving (32), (33) for the first steps. For  $k = 0$ , we require that  $\{F_0, S_0\} = 0$ . With our matrix notation,

$$\{F_0, S_0\} \equiv \{Ax, Bx\} = -[A, B]_{ij}x^j \quad (37)$$

and therefore at this stage we only have to determine the matrices  $B$  commuting with a given matrix  $A$ .

For  $k = 1$ , we just get  $\{F_0, S_1\} + \{F_1, S_0\} = 0$ ; in the matrix notation this just reads  $\{Ax, S_1\} = \{Bx, F_1\}$  or, using the homological operators notation,

$$\mathcal{A}(S_1) = \mathcal{B}(F_1) := \Psi_1(x). \quad (38)$$

For  $k = 2$  we get in the same way

$$\mathcal{A}(S_2) = \mathcal{B}(F_2) - \{F_1, S_1\} := \Psi_2(x), \quad (39)$$

and so on for any  $k$ .

**Remark 3** The fact that we can proceed recursively (and only deal with linear PDEs) does not mean that we are guaranteed to find solutions at any stage  $k$ . At  $k = 0$  we always have at least the solutions given by  $B = I$  and by  $B = A^q$  ( $q = 1, 2, \dots$ ). For  $k \geq 1$  we do not, in general, have solutions to the determining equations apart from  $S_j = cF_j$  for all  $q = 0, \dots, k$  (i. e., as  $k$  is generic,  $Y = cX$ ). This corresponds to the fact that symmetry is not generic.

**Remark 4** As the relevant equations are not homogeneous,  $S_k = 0$  is not, in general, an acceptable solution at stage  $k$ . This is quite natural if one thinks that the choice  $B = I$  is always acceptable at  $k = 0$ . Choosing  $S_k = 0$  at all the following stages would leave us with the dilation vector field  $Y = x^i \partial_i$ , which is a symmetry only for  $X$  linear [80,151,174].

### Approximate Symmetries

Note that it could happen where we are only able to determine a commuting vector field for  $X$  up to some finite order  $k$  (either for a full symmetry does not exist or for our limited capacities, computational or otherwise).

If in this case we consider a neighborhood of the origin of small size, say a ball  $\mathcal{B}_\varepsilon$  of radius  $\varepsilon \ll 1$ , in this we have  $[X, Y] = O(\varepsilon^k)$ ; thus,  $Y$  represents an *approximate symmetry* for  $X$ .

Approximate symmetries are interesting and useful in a number of contexts. In particular, in some cases – notably, for Hamiltonian vector fields – there is a connection between symmetries of dynamical systems and conserved quantities (i. e. constants of motion) for it; in this case, approximate symmetries will correspond to approximate constants of motion, i. e. to quantities which are not exactly conserved, but are approximately so.

More precisely, an approximate symmetry will correspond to a quantity  $J$  whose evolution under the dynamics described by  $X$  is slow of order  $k$ , i. e.  $dJ/dt \approx \varepsilon^k$  for some finite  $k$ . It is rather clear that these can be quite useful in applications, where we are often concerned with study of the dynamics over finite times; see [51] and especially, in the Hamiltonian case, [100].

### Symmetry Characterization of Normal Forms

Let us now consider the case where the dynamical system (13) has already been taken into Poincaré–Dulac normal form. We start by recalling some notions of linear algebra of use here.

#### Linear Algebra

A real matrix  $T$  is *semisimple* if its complexification  $T_C$  can be diagonalized, and is *normal* if it commutes with its adjoint,  $[T, T^+] = 0$ . A diagonal matrix is normal. For a normal matrix,  $T: \text{Ker}(T^+) \rightarrow \text{Ker}(T^+)$ . If  $T$  is normal we actually have  $\text{Ker}(T) = \text{Ker}(T^+)$ .

Any semisimple matrix can be transformed into a normal one by a linear transformation. If two semisimple matrices  $A, B$  commute, then they can be simultaneously diagonalized (by a linear, in general non-orthogonal, transformation), and so taken simultaneously to be normal. Thus, when considering such a pair of matrices, we can with no loss of generality assume them to be diagonal or, a fortiori, normal.

If we want to transform  $T$  into a real normal matrix, we just have to consider the transformation of  $T$  into a block diagonal matrix, the blocks corresponding to (complex

conjugate) eigenvalues. It is easy to see that in this way we still get a (real) normal matrix.<sup>(8)</sup>

In the following, we will at several points restrict, for ease of discussion, to normal matrices; our statements for normal matrices will be easily extended to semisimple ones up to the appropriate linear transformation.

#### Normal Forms

We now note that  $F \in \text{Ker}(\mathcal{A})$  means that the vector field associated with  $F$  (which we denote by  $X_F$ ) commutes with the linear vector field  $X_A$  associated with  $A$ . That is,

$$F \in \text{Ker}(\mathcal{A}) \Leftrightarrow \{Ax, F(x)\} = 0 \Leftrightarrow [X_A, X_F] = 0. \quad (40)$$

Thus we have the following characterization for vector fields in normal form (note this uses Lemma 1 and hence the scalar product defined in Sect. “[Perturbation Theory: Normal Forms](#)”).

**Lemma 3** *A vector field  $X = (A_{ij}x^j + F^i(x))\partial_i$ , where the  $F$  are nonlinear functions, is in normal form if and only if its nonlinear part  $X_F = F^i(x)\partial_i$  commutes with the vector field  $X_{A^+} = [A_{ij}^+x^j]\partial_i$  associated with the adjoint of its linear part, i. e. if and only if  $F \in \text{Ker}(\mathcal{A}^+)$ .*

For the sake of simplicity, we will only consider the case where the matrix  $A$ , corresponding to the linear part of  $X$  (in both the original and the normal form coordinates), commutes with its adjoint, i. e. we make the following<sup>(9)</sup> **assumption:** *The matrix  $A$  is normal:*  $[A, A^+] = 0$ .

If  $A$  is normal, then it follows  $[L_A, L_{A^+}] = 0$  due to Lemma 2; this implies in particular that:

**Lemma 4** *If  $A$  is normal, then  $\text{Ker}(\mathcal{A}) = \text{Ker}(\mathcal{A}^+)$ .*

It is important to recall that with the standard scalar product, we have  $(L_A)^+ = L_{A^+}$ . It is also important, although trivial, to note that if  $A$  is normal, then  $\mathcal{A}$  is a normal operator (under the standard scalar product), and  $\text{Ker}(\mathcal{A}) \cap \text{Ran}(\mathcal{A}) = \{0\}$ , but  $f \in \text{Ker}(\mathcal{A}^2) \Rightarrow (\mathcal{A}(f)) \in \text{Ker}(\mathcal{A})$  and therefore  $\mathcal{A}(f) = 0$ . Hence

**Lemma 5** *If  $A$  is normal,  $\text{Ker}(L_A^2) = \text{Ker}(L_A)$ .*

This discussion leads to a natural characterization of Poincaré–Dulac normal forms in terms of symmetry properties.<sup>(10)</sup>

**Lemma 6** *If  $A$  is a normal matrix, then  $X = (Ax + F)^i(\partial/\partial x^i)$  is in normal form if and only if  $X_A$  is a symmetry of  $X$ , i. e.  $F \in \text{Ker}(\mathcal{A})$ .*



### The General Case

In the general case, i.e. when  $A$  is not normal, the resonant terms will be those in  $\text{Ker}(\mathcal{A}^+)$ ; similarly, in this case we could characterize systems in normal form by  $\{F(x), A^+x\} = 0$ .

However, for a symmetry characterization it is better to proceed in a slightly different way. That is, we recall that any matrix  $A$  can be uniquely decomposed as  $A = A_s + A_n$  where  $A_s$  is semisimple and  $A_n$  is nilpotent, with  $[A_s, A_n] = 0$ . Resonance properties involve eigenvalues of  $A_s$ , and resonant terms will satisfy  $\{F(x), A_s^+x\} = 0$ .

This is a more convenient characterization, in that it shows that the full vector field (in normal form)  $X$  will commute with the linear vector field  $S = (A_s)^i_j x^j \partial_i$  corresponding to the semisimple part of its linear part.

### Symmetries and Transformation to Normal Form

We want to consider the case where the dynamical system (1), or equivalently the vector field  $X$ , admits a symmetry  $Y$  (the case of an  $n$ -dimensional symmetry algebra will be considered later on); we want to discuss how the presence of the symmetry affects the normalization procedure. Moreover, as the dynamical and symmetry vector fields are on equal footing, it will be natural to investigate if they can be both put into normal form; or even if some kind of **joint normal form** is possible (as is the case).

We will use the notation introduced above for the expression of  $X, Y$  in the  $x$  coordinates, and denote by  $y$  the normal form coordinates. Correspondingly, the bracket  $\{.,.\}$  (which is defined in coordinates) will be denoted as  $\{.,.\}_{(x)}$  or  $\{.,.\}_{(y)}$  when confusion is possible. We have, therefore,

$$\begin{aligned} X &= f^i(x)(\partial/\partial x^i) = g^i(y)(\partial/\partial y^i), \\ Y &= s^i(x)(\partial/\partial x^i) = r^i(y)(\partial/\partial y^i); \end{aligned} \quad (41)$$

and similarly for the power series expansions of  $f, g, s, r$  in terms homogeneous of degree  $(k+1) = 1, 2, \dots$ . We will denote the matrices associated with the linear parts of  $X, Y$  by, respectively,  $A$  and  $B$ :  $(Df)(0) = (Dg)(0) = A$ ,  $(Ds)(0) = (Dr)(0) = B$ . The corresponding homological operators will be denoted by  $\mathcal{A} = L_A$  and by  $\mathcal{B} = L_B$ . We assume that both  $A$  and  $B$  are normal matrices.

### Nonlinear Symmetries (The General Case)

The key, albeit trivial, observation is that the geometric relation  $[X, Y] = 0$  does not depend on the coordinate system we are using. Therefore, if  $\{f, s\}_{(x)} = 0$ , we must also have  $\{g, r\}_{(y)} = 0$ .

Another important, and again quite trivial, observation is that when we consider a P-transformation  $x = y + h_k(y)$ , the term  $F_k$  of order  $k$  in  $X$  changes according to  $\mathcal{A}(h_k)$ , but the term  $S_k$  of order  $k$  in  $Y$  changes according to  $\mathcal{B}(h_k)$ . The same applies when we consider a Lie-Poincaré transformation.

Thus, although when we choose  $h_k + \ell_k$  [with  $\ell_k \in \text{Ker}(\mathcal{A})$ ] to generate the Poincaré transformation, we get the same transformation as that generated by  $h_k$  on the  $F_k$  (see Remark 1), the transformation on  $S_k$  can be different. This means that the freedom left by the Poincaré prescription for construction of the normalizing transformation could, in principle, be used to take the symmetry vector field  $Y$  into some convenient form. This is indeed the case, as will be shown below.

Two vector fields  $X, Y$  on  $M$ , as in (35), with  $A$  and  $B$  semisimple, are in **Joint Normal Form** if both  $G_k$  and  $R_k$  are in  $\text{Ker}(\mathcal{A}) \cap \text{Ker}(\mathcal{B})$  for all  $k \geq 1$ .

**Theorem 3** *Let the vector fields  $X = f^i(x)\partial_i$  and  $Y = s^i(x)\partial_i$  have a fixed point in  $x_0 = 0$ . Let them commute,  $[X, Y] = 0$ , and have normal semisimple linear parts  $A = (Df)(0)$  and  $B = (Ds)(0)$ . Then, by means of a sequence of Poincaré transformations, they can be brought to Joint Normal Form.*

In this theorem (a proof of this is given in [49,51]; see also [106] for a different approach to a related problem) we did not really use the interpretation of one of the vector fields as describing the dynamics of the system and the other describing a symmetry, but only their commutation relation. From this point of view, it is natural to also consider arbitrary (possibly non-Abelian) algebras of vector fields.

### Linear Symmetries

A special case of symmetries is given by *linear* symmetries, i.e. by the case where

$$Y = X_B = \left( B_{ij} x^j \right) (\partial/\partial x^i). \quad (42)$$

In this case, if  $X, Y$  are in Joint Normal Form we have in particular that  $f \in \text{Ker}(\mathcal{B})$ . We have the following corollaries to Theorem 3:

**Corollary 1** *If the linear vector field  $Y = X_B$  is a symmetry for  $X = f^i(x)\partial_i$ , it is possible to normalize  $f$  by passing to  $y$  coordinates so that  $X = g^i(y)(\partial/\partial y^i)$ ,  $Y = (By)^i(\partial/\partial y^i)$ , and  $g \in \text{Ker}(\mathcal{A}) \cap \text{Ker}(\mathcal{B})$ .*

**Corollary 2** *Let  $s(x) = Bx + S(x)$ , with  $S$  the nonlinear part of  $s$ , and let  $Y = s^i(x)\partial_i$  be a symmetry of*

$X = f^i(x)\partial_i$ . Then when  $X, Y$  are put in Joint Normal Form,  $X_B$  is a linear symmetry of  $X$ .

When we perform the Poincaré transformations needed to transform  $X$  into its normal form, there seems to be no reason, a priori, why the  $Y$  should keep its linear form. It is actually possible to prove the following result (the proof of this relies on a similar result by Ruelle [159,160] dealing with the center manifold mapping, and is given e.g. in [60,118]; see also [20,21,22,23]) that we quote from [118]:

**Theorem 4** *If  $X$  commutes with a linear vector field  $Y = X_B = (B_{ij}x^j)\partial_i$ , then it is possible to find a normalizing series of Poincaré transformations with generators  $h_k \in \text{Ker}(B)$ , so that in the new coordinates  $y$ ,  $X$  is taken into normal form and  $Y$  is left unchanged, i.e.  $Y = (B_{ij}y^j)\partial_i$ .*

Note that, for resonant  $B$ , Theorem 4 is *not* a special case of Theorem 3 and the above Corollaries.<sup>(11)</sup>

**Remark 5** One should avoid confusions between linear symmetries of the dynamical system and symmetries of its linearization, which do not extend in general to symmetries of the full system.

## Generalizations

In this section we are going to discuss some generalizations of the results illustrated in the previous one: we will deal with the problem of transformation into normal form of a (possibly non Abelian) Lie algebra with more than two generators, not necessarily commuting.<sup>(12)</sup>

### Abelian Lie Algebra

It is actually convenient to drop the distinction between the vector field defining the dynamical system and the symmetry vector fields. Thus, we simply consider an algebra  $\mathcal{G}$  of vector fields  $X_i$ .<sup>(13)</sup>

First of all, we consider the case of an Abelian Lie algebra of vector fields. In this case we have the following result (see [51] for a proof).

**Theorem 5** *Let  $\{X_1, \dots, X_r\}$  commute, and assume that the matrices  $A_{(i)}$  identifying the linear parts of the  $X_j$  are normal. Then  $\{X_1, \dots, X_r\}$  can be put in Joint Normal Form by a sequence of Poincaré or Lie-Poincaré transformations.*

Note that if  $\mathcal{A}_i$  are the homological operators corresponding to the  $A_{(i)}$ , and we write  $\mathcal{K} = \bigcap_{i=0}^k \text{Ker}(\mathcal{A}_i)$ , the Theorem states there are coordinates  $y^i$  such that  $X_j = (\tilde{f}_{(j)}(y))^i (\partial/\partial y^i)$  with  $\tilde{f}_{(j)} \in \mathcal{K}$  for all  $j$ .

### Nilpotent Lie Algebra

For generic Lie algebras one cannot expect results as general as in the Abelian case [4,48,51]. A significant exception to this is met in the case of *nilpotent* algebras<sup>(14)</sup> (see also [54] for a group-theoretical approach), in which we can recover essentially the same results obtained in the Abelian case (see [112] for the case of semisimple Lie algebras).

Actually, an extension of Theorem 5 to the nilpotent case should be considered with some care, as the only nilpotent algebras of nontrivial semisimple matrices are Abelian. On the other hand, we could have a non-Abelian nilpotent algebra of vector fields with linear parts given by semisimple matrices, provided that some of these vanish.

Indeed, although  $[X_i, X_j] = c_{ij}^k X_k$  necessarily implies that  $[A_i, A_j] = c_{ij}^k A_k$ , it could happen that all the  $A_k$  for which there exists a nonzero  $c_{ij}^k$  do vanish, so that the algebra of vector fields is “Abelian at the linear level”. (As a concrete example, consider the algebra spanned by  $X = -x^2 d/dx$  and  $Y = x d/dx$ .)

Note that in this case  $\text{Ker}(\mathcal{A}_k)$  would be just the whole space; needless to say, we should consider the full vector fields  $X_k$ , which will produce (by assumption) a closed Lie algebra.

With this remark (and in view of the fact that the proof of Theorem 5 is based on properties of the algebra of the  $A_i$ ’s and of the corresponding homological operators, see [51]) it is to be expected that the result for the nilpotent case will not substantially differ from the one holding for the Abelian case (as usual, the key to this extension will be to proceed to normalization of vector fields in an order which respects the structure of the Lie algebra). This is indeed what happens, and one has the following result [48,51]:

**Theorem 6** *Let the vector fields  $\{X_1, \dots, X_r\}$  form a nilpotent Lie algebra  $\mathcal{G}$  under  $[\cdot, \cdot]$ ; assume that the matrices  $A_{(i)}$  identifying the linear parts of the  $X_j$  are normal. Then  $\{X_1, \dots, X_r\}$  can be put in Joint Normal Form by a sequence of Poincaré or Lie-Poincaré transformations.*

**Corollary 3** *If the general Lie algebra  $\mathcal{G}$  of vector fields  $\{X_1, \dots, X_n\}$  contains a nilpotent subalgebra  $\mathcal{G}^*$ , then the set of vector fields  $X_i$  spanning this  $\mathcal{G}^*$  can be put in Joint Normal Form.*

### General Lie Algebra

Some “partial” Joint Normal Form can be obtained, even for non-nilpotent algebras, under some special assumptions. We will just quote a result in this direction, referring as usual to [51] for details. A description of normal

forms for systems with symmetry corresponding to simple compact Lie groups is given in [86]

**Theorem 7** *Let  $\mathcal{G}$  be a  $d$ -dimensional algebra spanned by  $X_{F_a}$  with  $F_a = A_a x + F_a(a = 1, \dots, d)$ , and let  $\mathcal{G}$  admit a non-trivial center  $C(\mathcal{G})$ . Let the center of  $\mathcal{G}$  be spanned by  $X_{w_b}$  with  $w_b = C_b x + W_b(b = 1, \dots, d_C)$ , where  $d_C \leq d$ , and assume that the semisimple parts  $C_{b,s}$  are normal matrices. Denote by  $C_{b,s}$  the associated homological operators, and write  $\mathcal{K}_s = \cap_{b=1}^{d_C} \text{Ker}(C_{b,s})$ .*

*Then, by means of a sequence of Poincaré transformations, all the  $F_a$  can be taken into  $\hat{F}_a \in \mathcal{K}_s$ . The same holds for  $\hat{\mathcal{G}} = \mathcal{G} \oplus \mathcal{N}$ , with  $\mathcal{N}$  a nilpotent subalgebra.*

### Symmetry for Systems in Normal Form

No definite relation exists, in general, between symmetries  $\mathcal{G}_X$  of a vector field  $X$  and symmetries  $\mathcal{G}_A$  of its linear part, or between constants of motion  $\mathcal{I}_X$  for  $X$  and constants of motion  $\mathcal{I}_A$  for its linear part, but if  $X$  is in normal form, one has some interesting results [48,186]:

**Lemma 7** *If  $X$  is in normal form, any constant of motion of  $X$  must also be a constant of motion of its linearization  $A$ , i. e.  $\mathcal{I}_X \subseteq \mathcal{I}_A$ .*

In general  $\mathcal{I}_X \neq \mathcal{I}_A$ , even if  $X$  is in normal form. Also, if  $[B, A] = 0$  in general  $\mu(x)Bx$  does not belong to  $\mathcal{G}_X$ , even for  $\mu \in \mathcal{I}_X$  (unless  $X_B \in \mathcal{L}_X$  as well, where  $\mathcal{L}_X$  are the linear symmetries of  $X$ ), nor to  $\mathcal{G}_A$  (unless  $\mu \in \mathcal{I}_A$ ).

**Lemma 8** *If  $X$  is in normal form, then  $\mathcal{G}_X \subseteq \mathcal{G}_A$ .*

This result allows for the restricting our search for  $Y \in \mathcal{G}_X$  to  $\mathcal{G}_A$  rather than considering the full set of vector fields on  $M$ . Similarly, Lemmas 7 and 8 can be useful in the determination of the sets  $\mathcal{L}_X, \mathcal{I}_X$ , in that we can first solve the easier problem of determining  $\mathcal{L}_A, \mathcal{I}_A$ , and then look for  $\mathcal{L}_X, \mathcal{I}_X$  in the class of vector fields  $\mathcal{L}_A$  and of the functions in  $\mathcal{I}_A$ , rather than in the full set of linear vector fields on  $M$  and, respectively, in the full set of scalar functions on  $M$ .

Moreover,  $\mathcal{G}_A$  can be determined in a relatively simple way, by solving the system of quasi-linear non-homogeneous first order PDEs  $\{Ax, g\} = 0$ , which are written explicitly as

$$(A_{ij}x^j)(\partial g^k / \partial x^i) = A_{kj}g^j. \quad (43)$$

By considering this, and introducing the set  $\mathcal{I}_A^*$  of the meromorphic (i. e., quotients of formal power series) constants of motion of the linear problem  $\dot{x} = Ax$ , one can obtain [48,60,186] the following result:

**Lemma 9**  *$\mathcal{G}_A$  is the set of all formal power series in  $\mathcal{I}_A^* \otimes \mathcal{L}_A$ .*

In a more explicit form, as  $\mathcal{G}_A \equiv \text{Ker}(\mathcal{A})$ , the resonant terms  $F(x) \in \text{Ker}(\mathcal{A})$  are power series of the form  $F(x) = K(\rho(x)) \cdot x$ , where  $K$  is a matrix commuting with  $A$  when written in terms of its real entries  $K_{ij}$ , and where  $K(\rho(x))$  is the same matrix in which the entries  $K_{ij}$  are replaced by functions of the constants of motion  $\rho = \rho(x) \in \mathcal{I}_A^*$ . The set of the vector fields in  $\mathcal{G}_A$  is of course a Lie algebra.

We summarize our discussion for dynamical systems in normal form in the following proposition:

**Theorem 8** *Let  $X$  be a vector field in normal form, and let  $A$  be the normal matrix corresponding to its linear part. Then,  $\mathcal{L}_X \subseteq \mathcal{G}_X \subseteq \mathcal{G}_A$ ; and  $\mathcal{L}_X \subseteq \mathcal{L}_A \subseteq \mathcal{G}_A$ .*

**Remark 6** It should be mentioned that Kodama considered the problem of determining  $\mathcal{G}_A$  from a more algebraic standpoint [122]. In the same work, Kodama also observed that  $\mathcal{G}_A$ , considered as an algebra, is not only infinite-dimensional, but has the natural structure of a graded Virasoro algebra.

**Remark 7** We emphasize once again that the above results were given for  $X$  in normal form. They can obviously be no longer true if  $X$  is not in normal form.<sup>(15)</sup>

### Linearization of a Dynamical System

An interesting application of the Joint Normal Form deals with the case of *linearizable* dynamical systems. Clearly, if  $\text{Ker}(\mathcal{A}) = \{0\}$ , the dynamical system is linearizable by means of a formal Poincaré transformation. But, whatever the matrix  $A$ , the linear vector field  $X_A = (Ax \cdot \nabla)$  commutes with the vector field  $S = \sum_i x_i(\partial/\partial x_i) = ((Ix) \cdot \nabla)$ , which generates the dilations in  $\mathbf{R}^n$ . It is easy to see that, conversely, the only vector fields commuting with  $S$  are the linear ones. It is also clear that the identity does not admit resonances. Thus [16,92]:

**Lemma 10** *A vector field  $X_f$  (or a dynamical system  $\dot{x} = f(x)$ ) can be linearized if and only if it admits a (possibly formal) symmetry  $X_g$  such that  $B = (Dg)(0) = I$ .*

Proceeding in a similar way – but using Joint Normal Forms – we have, more generally:

**Theorem 9** *The vector field  $X_f$  with  $A = (Df)(0)$  can be linearized if and only if it admits a (possibly formal) symmetry  $X_g$  with  $B = (Dg)(0)$  such that  $A$  and  $B$  do not admit common resonances, i. e.  $\text{Ker}(\mathcal{A}) \cap \text{Ker}(\mathcal{B}) = \{0\}$ .*

This result can be easily extended not only to the case of more than one symmetry, as an obvious consequence of Theorem 4, but also to the non-semisimple case [48].

Another interesting result related to linear dynamical systems, is the following [48,51]:

**Theorem 10** *If a dynamical system in  $\mathbf{R}^n$  can be linearized, then it admits  $n$  independent commuting symmetries, that can be simultaneously linearized.*

### Further Normalization and Symmetry

As repeatedly noted above (see in particular Remark 1), when the linear part of the dynamical vector field is resonant the resulting degeneracy in the solution to the homological equation is not a real degeneracy for what concerns effects on higher order terms in the normal form.

These higher order terms could – and in general will – generate resonant terms, which cannot be eliminated by the standard algorithm. On the other hand, it is clear that this could be seen as a bonus rather than as a problem: in fact, it is conceivable that by carefully choosing the component of  $h_k$  lying in  $\text{Ker}(\mathcal{A})$ , one could generate resonant terms which exactly cancel those already present in the vector field.

Several algorithms have been designed to take advantage, in one way or the other, of this possibility; some review of different approaches is provided in the paper [39]. Here we are concerned with those based on symmetry properties, and discuss two different approaches, developed respectively by the present author [81,83] and by Palacian and Yanguas [155].

It should be stressed that once the presence of additional symmetries – and in the Hamiltonian context, additional constants of motion – has been determined for the normal form truncated at some finite order  $N$ , one should investigate if the set of symmetries (or constants of motion) persist under small perturbations; in particular, when considering terms of higher order as well. A general tool to investigate this kind of question is provided by the Nekhoroshev generalization of the Poincaré-Lyapounov theorem [148,149,150]; see also the discussion in [15,87,88,91].

### Further Normalization and Resonant Further Symmetry

We will assume again, for ease of discussion, that the matrix  $A$  associated with the linear part of the dynamical vector field  $X$  is normal. In this case, as discussed above, the normal form is written as  $X = g^i(x)\partial_i$  where  $g(x) = \sum_{k=0}^{\infty} G_k(x)$ , with  $G_k \in \mathcal{V}_k$  and all  $G_k$  being reso-

nant. We can correspondingly write

$$X = \sum_{k=0}^{\infty} X_k, X_k = G_k^i \partial_i. \quad (44)$$

As  $X$  is in normal forms, we are guaranteed to have

$$[X_0, X_k] = 0 \quad \forall k = 0, 1, 2, \dots; \quad (45)$$

this corresponds to the characterization of normal forms in terms of symmetry as discussed in Sect. “Symmetry Characterization of Normal Forms”. In other words,  $G_k \in \text{Ker}(\mathcal{A})$  and hence, defining  $\mathcal{G}_0$  as the Lie algebra of vector fields commuting with  $X_0$ ,  $X_k \in \mathcal{G}_0$ . On the other hand, in general it will be  $[X_j, X_k] \neq 0$  for  $j \neq k$  and both  $j, k$  greater than zero; thus  $\mathcal{G}_0$  is not, in general, an Abelian Lie algebra: we can only state that  $X_0$  belongs to the center of  $\mathcal{G}_0$ ,  $X_0 \in Z(\mathcal{G}_0)$ .

Suppose we want to operate further Lie-Poincaré transformations generated by functions which are symmetric under  $X_0$  (i. e. are in the kernel of  $\mathcal{A}$ ). It follows from the formulas obtained in Sect. “Perturbation Theory: Normal Forms” that these will map  $\mathcal{G}_0$  into itself, i. e.  $\mathcal{G}_0$  is globally invariant under this restricted set of Lie-Poincaré transformations.

As for the individual vector fields, it follows from the general formula (24) that each of them is invariant under such a transformation *at first order in  $h_k$* , but not at higher orders. That is, making use again of Remark 1, we can still in this way generate new resonant terms in the normal form, including maybe terms which cancel some of those present in (44).

By looking at the explicit formulas (25), it is rather easy to analyze in detail the higher order terms generated in the concerned Lie-Poincaré transformation.

Note that we did not take into account problems connected with the convergence of the further normalizing transformations; it has to be expected that each step will reduce the radius of convergence, so that the further normalized forms will be actually (and not just formally) conjugated to the original dynamic in smaller and smaller neighborhoods of the fixed point; we refer to [90] for an illustration of this point by explicit examples and numerical computations; and to ► [Perturbative Expansions, Convergence of](#) for a general discussion on the convergence of normalizing transformations.

In order to state exactly the result obtained by this construction, we need to introduce some function spaces, which require abstract definitions. With<sup>(16)</sup>  $\mathcal{L}_k(\cdot) := [X_k, \cdot]$ , we set

$$H^{(p)} := \text{Ker}(\mathcal{L}_0) \cap \dots \cap \text{Ker}(\mathcal{L}_{p-1}) \quad (46)$$

(note  $H^{(q)} \subseteq H^{(p)}$  for  $q > p$ ), and denote by  $\mathcal{M}_p$  the restriction of  $\mathcal{L}_p$  to  $H^{(p)}$ ; thus  $\text{Ker}(\mathcal{M}_p) = H^{(p+1)}$ . We define spaces  $F^{(p)}$  (with  $F^{(0)} = \mathcal{V}$ ) as

$$F^{(p)} := \text{Ker}(\mathcal{M}_0^+) \cap \dots \cap \text{Ker}(\mathcal{M}_{p-1}^+); \quad (47)$$

the adjoint should be meant in the sense of the scalar product introduced in Sect. “[Perturbation Theory: Normal Forms](#)”. We also write  $F_k^{(p)} = F^{(p)} \cap \mathcal{V}_k$ .

We will say that  $X$  is in **Poincaré renormalized form** up to order  $N$  if  $G_k \in F_k^{(k)}$  for all  $k \leq N$ .

**Theorem 11** *The vector field  $X$  can be formally taken into Poincaré renormalized form up to (any finite) order  $N$  by means of a (finite) sequence of Lie-Poincaré transformations.*

For a proof of this theorem, and a detailed description of the renormalizing procedure, see [51,81,83,85]; see [51,81] for the case where additional symmetries are present; an improved procedure, taking full advantage of the Lie algebraic structure of  $\mathcal{G}_0$ , is described in [84].

### Further Normalization and External Symmetry

A different approach to further reduction (simplification) of vector fields in normal form has been developed by Palacian and Yanguas [155] (and applied mainly in the context of Hamiltonian dynamics [154,156,157]), making use of a result by Meyer [144].

As discussed in the previous subsection we have  $[X_0, X_k] = 0$  for all  $k$ . Suppose now there is a linear (in the coordinates used for the decomposition (44)) vector field  $Y$  such that  $[X_0, Y] = 0$ . Then the Jacobi identity guarantees that

$$[X_0, [X_k, Y]] = 0 \quad \forall k. \quad (48)$$

We assume that  $Y$  also corresponds to a normal matrix  $B$ , so that the homological operator  $\mathcal{B}$  associated to it is also normal.

We can then proceed to further normalization as above, being guaranteed that – provided we choose  $h_k \in \text{Ker}(\mathcal{L}_0)$  – the resulting vector fields will not only still be in  $\mathcal{G}_0$ , but also still satisfy (48). One can use freedom in the choice of the generator  $h_k \in \text{Ker}(\mathcal{L}_0)$  for the further normalization in a different way than discussed above: that is, we can choose it so that  $[Y, X_k] = 0$ ; in other words, we will get  $\hat{X} \in \text{Ker}(\mathcal{A}) \cap \text{Ker}(\mathcal{B})$ .

Note the advantage of this: we do not have to worry about complicated matters related to relevant homologi-

cal operators acting between different spaces, as we only make use of the homological operator  $\mathcal{B}$  associated with the “external” symmetry linear vector field and thus mapping each  $\mathcal{V}_k$  into the same  $\mathcal{V}_k$ .

The result one can obtain in this way is the following (which we quote in a simplified setting for the sake of brevity; in particular the normality assumption can be relaxed) [155].

**Theorem 12** *Let  $X$  be in normal form; assume moreover  $X_0 = Ax$  and there is a normal matrix  $B$  such that  $[A, B] = 0$ ; denote by  $Y$  the associated vector field,  $Y = (Bx)^i \partial_i$ . Assume moreover for each resonant vector field  $R_k$  there is  $Q_k$  satisfying  $[Y, Q_k] = 0$  and such that  $\hat{R}_k = R_k + [X_0, Q_k]$  commutes with  $Y$ . Then  $X$  can be taken to a (different) normal form  $\hat{X}$  such that  $[Y, \hat{X}] = 0$ .*

Applications of this theorem, and more generally of this approach, are discussed e.g. in [154,155,156,157]. We stress that albeit the assumptions of this theorem are rather strong<sup>(17)</sup>, it points out to the fact that *there are* cases in which a symmetry – and in the Hamiltonian framework, an integral of motion – of the linear part can be extended to a symmetry of the full normal form.

### Symmetry Reduction of Symmetric Normal Forms

Symmetry reduction is a general – and powerful – approach to the study of nonlinear dynamical systems. (In the Hamiltonian case, this is also known as (Marsden–Weinstein) moment map [6,134,135].)

A general theory based on the geometry of group action has been developed by Louis Michel; this was originally motivated by the study of spontaneous symmetry breaking in high-energy physics [140,141] (see [136,137,138,139] for the simpler case where only stationary solutions are considered, and [142] for the full theory and applications; see also [2,82,85,89,93,166,167]).

A description of this would lead us too far away from the scope of this paper, but as this theory also applies to vector fields in normal form, we will briefly describe the results that can be obtained in this way; we will mainly follow [94] (see [95,97] for further detail, extensions, and a more abstract mathematical formulation).

As mentioned in Sect. “[Symmetry of Dynamical Systems](#)”, the Lie algebra of vector fields in normal form is infinite dimensional, but also has the structure of a Lie module over the algebra of constants of motion for the linear part  $X_0$  of the vector field (which remains the same under all the considered transformations).



Let us recall that vector monomial  $\mathbf{v}_{\mu,\alpha} := x^\mu \mathbf{e}_\alpha$  is *resonant* with  $A$  if

$$\begin{aligned} (\mu \cdot \lambda) &:= \sum_{i=1}^n \mu_i \lambda_i = \lambda_\alpha \quad \text{with } \mu_i \geq 0, \\ |\mu| &:= \sum_{i=1}^n \mu_i \geq 1; \end{aligned} \quad (49)$$

here  $\lambda_i$  are the eigenvalues of  $A$ , which we suppose to be semisimple, for the sake of simplicity (in the general case one would consider  $A_s$  rather than  $A$ ).

As mentioned in Sect. “[Perturbation Theory: Normal Forms](#)”, the relation  $(\mu \cdot \lambda) = \lambda_\alpha$  is said to be a *resonance relation* related to the eigenvalue  $\lambda_\alpha$ , and the integer  $|\mu|$  is said to be the *order* of the resonance. In the present context it is useful to include order one resonances in the definition (albeit the *trivial* order one resonances given by  $\lambda_\alpha = \lambda_\alpha$  are obviously of little interest).

Let us consider again the resonance Eq. (49). It is clear that if there are non-negative integers  $\sigma_i$  (some of them nonzero) such that

$$\sum_{i=1}^n \sigma_i \lambda_i = 0, \quad (50)$$

then we always have infinitely many resonances. In this case the monomial  $\varphi = x^\sigma$  will be called a *resonant scalar monomial*. It is an invariant of  $X_0$ , and any multi-index  $\mu$  with  $\mu_i = k\sigma_i + \delta_{i\alpha}$  provides a resonance relation  $(\mu \cdot \lambda) = \lambda_\alpha$  related to the eigenvalue  $\lambda_\alpha$ ; in other words, any monomial  $x^{k\sigma} x^\alpha = \varphi^k x^\alpha$  is resonant, and so is any vector  $\mathbf{v}_{k\sigma + \mathbf{e}_\alpha, \alpha}$ .

Therefore, we say that (50) identifies a *invariance relation*. The presence of invariance relations is the only way to have infinitely many resonances in a finite dimensional system (see [186]).

Any nontrivial resonance (49) which does not originate in an invariance relation, is said to be a *sporadic resonance*. Sporadic resonances are always in finite number (if any) in a finite dimensional system [186].

Any invariance relation (50) such that there is no  $v$  with  $v_i \leq \sigma_i$  (and of course  $v \neq \sigma$ ) providing another invariance relation, is said to be an *elementary invariance relation*. Every invariance relation is a linear combination (with nonnegative integer coefficients) of elementary ones. Elementary invariance relations are always in finite number (if any) in a finite dimensional system [186]. If there are  $m$  independent elementary invariance relations, each of them of the form (50), we associate to these monomials  $\beta_j = x^\sigma = \prod_{i=1}^n x_i^{\sigma_i}$  ( $j = 1, \dots, m$ ).

Similarly, if there are  $r$  sporadic resonances (49), we associate resonant monomials  $\alpha^j(x) = x^\mu = \prod_{i=1}^n x_i^{\mu_i}$  ( $j = 1, \dots, r$ ) and resonant vectors  $\mathbf{v}_{\mu,\alpha}^{(j)}$  to sporadic resonances. We then introduce two set of new coordinates: these will be the coordinates  $w^1, \dots, w^r$  in correspondence with sporadic resonances, and other new coordinates  $\varphi^1, \dots, \varphi^m$  in correspondence with elementary invariance relations. The evolution equations for the  $x^i$  can be written in simplified form using these (note that some ambiguity is present here, in that we can write these in different ways in terms of the  $x, w, \varphi$ ), but we should also assign evolution equations for them; these will be given in agreement with the dynamics itself. That is, we set

$$\frac{dw^j}{dt} = \frac{\partial w^j}{\partial x^i} \frac{dx^i}{dt} := h^j(x, w, \varphi); \quad (51)$$

$$\frac{d\varphi^a}{dt} = \frac{\partial \varphi^a}{\partial x^i} \frac{dx^i}{dt} := z^a(x, w, \varphi). \quad (52)$$

We are thus led to consider the enlarged space  $W = \mathbf{R}^{n+r+m}$  of the  $(x, w, \varphi)$ , and in this the vector field

$$\begin{aligned} Y &= f^i(x, w, \varphi) (\partial/\partial x^i) + h^j(x, w, \varphi) (\partial/\partial w^j) \\ &\quad + z^a(x, w, \varphi) (\partial/\partial \varphi^a). \end{aligned} \quad (53)$$

The vector field  $Y$  is uniquely defined on the manifold identified by  $\psi^j := w^j - \alpha^j(x) = 0, \varphi^a - \beta^a(x) = 0$ . It is obvious (by construction) that the  $(n+m)$ -dimensional manifold  $M \subset W$  identified by  $\psi^i := w^i - \alpha^{(i)} = 0$  is invariant under the flow of  $Y$ , see (51). It is also easy to show that the functions  $z^a$  defined in (52) can be written in terms of the  $\varphi$  variables alone, i. e.  $\partial z^a / \partial x^i = \partial z^a / \partial w^j = 0$ . This implies<sup>(18)</sup>

**Lemma 11** *The evolution of the  $\varphi$  variables is described by a (generally, nonlinear) equation involving the  $\varphi$  variables alone.*

Note that the equations for  $x$  and  $w$  depend on  $\varphi$  and are therefore non-autonomous. We have the following result (we refer to [94,95] for a proof; see [97] for extensions).

**Theorem 13** *The analytic functions  $f^i$  and  $h^j$  defined above can be written as linear in the  $x$  and  $w$  variables, the coefficients being functions of the  $\varphi$  variables. Hence the evolution of the  $x$  and  $w$  variables is described by non-autonomous linear equations, obtained by inserting the solution  $\varphi = \varphi(t)$  of the equations for  $\varphi$  in the equations  $\dot{x} = f(x, w, \varphi), \dot{w} = h(x, w, \varphi)$ .*

Note that if no invariance relations are present, hence no  $\varphi$  variables are introduced, then the system describing the time evolution of the  $x, w$  variables is linear; in this case we can interpret normal forms as projections of a linear

system to an invariant manifold (without symmetry reduction).

If there are no sporadic resonances of order greater than one, then upon solving the reduced equation for the  $\varphi$  variables one obtains a non-autonomous linear system. Moreover, if all eigenvalues are distinct then we have a product system of one-dimensional equations.

If  $\varphi(t)$  converges to some constant  $\varphi_0$ , the asymptotic evolution of the system is governed by a linear autonomous equation for  $x$  and  $w$ . Similarly, if there is a periodic solution  $\bar{\varphi}(t)$  and  $\varphi(t)$  converges to  $\bar{\varphi}(t)$  for large  $t$ , the asymptotic evolution of the system is governed by a linear equation with  $t$ -periodic coefficients for  $x$  and  $w$ .

## Conclusions

We have reviewed the basic notions concerning symmetry of dynamical systems and its determination, in particular in a perturbative setting. We have subsequently considered various situations where the interplay of perturbation theory and symmetry properties produce nontrivial results, either in that the perturbative expansion turns out to be simplified (with respect to the general case) due to symmetry properties, or in that computations of symmetry is simplified by dealing with a system in normal form; we then considered the problem of jointly normalizing an algebra of vector fields (with possibly but not necessarily one of these defining a dynamical system, the other being symmetry vector fields). We also discussed how normal forms can be characterized in terms of symmetry, and how this is extended to a characterization of “renormalized forms”. Finally we considered symmetry reduction applied to systems in normal form.

The discussion conducted here illustrates some of the powerful conceptual and computational simplifications arising for systems with symmetry, also in the realm of perturbation theory.

As remarked in the Introduction, symmetry is a non-generic property; on the other hand, it is often the case that equations arising from physical (mechanical, electronic, etc.) systems enjoy some degree of symmetry, as a consequence of the symmetry of the fundamental equations of physics. Disregarding the symmetry properties in these cases would mean renouncing the use of what is often the only handle to grab the behavior of non-linear systems; and correspondingly a symmetry analysis can often on the one hand lead to identifying several relevant properties of the system even without a complete solution, and on the other hand being instrumental in obtaining exact (or approximate, as we are here dealing with perturbation theory) solutions.

Here we discussed some of the consequences of symmetry for the specific case of dynamical systems, such as those met in analyzing the behavior of nonlinear systems near a known (e. g. trivial) solution.

For a more general discussion, as well as for concrete applications, the reader is referred on the one hand to texts discussing symmetry for differential equations [3,19,26,27,37,80,115,125,151,152,174], on the other to texts and papers specifically dealing with the interplay of symmetry and perturbation theory, quoted in the main text and listed in the ample Bibliography below.

## Future Developments

First and foremost, future developments should be expected to concern further application of the general theory in concrete cases, both in nonlinear theoretical mechanics and in specific subfields, ranging from the more applied (e. g., ship dynamics and stability [13,176]; or handling of complex electrical networks [133,183]) to the more theoretical (e. g., galactic dynamics [24,59]).

On the other hand, the theory developed so far is in many cases purely formal, in that consideration of convergence properties – and estimation of the convergence region in phase and/or parameter space – of the resulting (perturbative) series is often left aside, with the understanding that in any concrete application one will have explicit series and be ready to analyze their convergence. The story of general (i. e. non-symmetric) perturbation theory shows however that the theoretical analysis of convergence properties can be precious – not only for the conceptual understanding but also in view of concrete applications – and it should thus be expected that future developments will deal with convergence properties in the symmetric case (see e. g. ► [Perturbative Expansions, Convergence of](#), [52] for a review of existing results).

The same issue of convergence, and estimation of the convergence region, arises in connection with further normalization (under any approach), and has so far been given little consideration.

A different kind of generalization is called for when dealing with symmetry reduction of normal forms: in fact, on the one hand it is natural to try applying the same approach (based on quite general geometrical properties) to more general systems than initially considered; on the other hand the method discussed in Sect. “[Symmetry Reduction of Symmetric Normal Forms](#)” is algorithmic and could be implemented by symbolic manipulations packages – such as those already existing for computations of symmetry of differential equations – which would be of

help to anybody having to deal with perturbation of concrete symmetric systems.

Finally, another field of future developments can be called for: here we discussed the interplay of perturbation theory and “standard” symmetries. Or, the notion of symmetry of differential equations has been generalized in various directions, producing in some cases a significant advantage in application to concrete systems. It should thus be expected that the interplay between these generalized notions of symmetry and perturbation theory will be investigated in the near future, and most probably will produce interesting and readily applicable results.

### Additional Notes

We collect here the additional notes called throughout the manuscript.

- (1) Note that  $\widetilde{M}$  is naturally a *fiber bundle* [40,119,146,147] over  $t$ : that is, it can be decomposed as the union of copies of  $M$ , each one in correspondence to a value of  $t$ . A *section* of this bundle is simply the graph of a function  $\sigma : t \rightarrow M$ . The set  $\sigma_x$  considered in a moment is a section of this fiber bundle.
- (2) For general differential equations, one would go along the same lines. A relevant difference is however present: for (systems of) first order equations there is no algorithmic way to find the general solutions to determining equations, as opposed to any other case [151,174].
- (3) In the case of Hamiltonian systems, one can work directly on the Hamiltonian  $H(p, q)$  rather than on the Hamilton equations of motion ► [Hamiltonian Perturbation Theory \(and Transition to Chaos\)](#), [111]. We will however, for the sake of brevity, not discuss the specific features of the Hamiltonian case.
- (4) Later on, in particular when we deal with different homological operators, it will be convenient to also denote this  $\mathcal{A}$  as  $L_A$ , with reference to the matrix  $A$  appearing in  $F_0 = Ax$ .
- (5) This is equivalent to defining  $(\xi_{\mu,i}, \xi_{\nu,j}) = \delta_{\mu,\nu} \delta_{i,j} (\mu!)$ , where for the multi-index  $\mu$  we have defined  $\mu! = (\mu_1!) \dots (\mu_n!)$ .
- (6) The name “resonant” is due to the relation existing between eigenvectors of  $\mathcal{A}$  and resonance relations among eigenvalues of the matrix  $A = (Df)(0)$  describing the linear part  $F_0(x) = Ax$  of the system.
- (7) In order to determine  $\widetilde{X} = \widetilde{f}^i(x) \partial / \partial x^i$ , we write (using the exponential notation)  $\xi(t + \delta t) = e^{\delta t X} \xi(t)$ . Therefore,  $x(t + \delta t) = [e^{sH} e^{(\delta t)X} \xi(t)]_{s=1}$ . Using  $\xi(t) = e^{-H} x(t)$ , we have  $[x(t + \delta t) - x(t)] = [e^{sH} (e^{(\delta t)X} - I) e^{-sH} x(t)]_{(s=1)}$ , and therefore (24).
- (8) Note, however, that when the diagonalizing matrix  $M$  is not unitary, this transformation changes implicitly the scalar product.
- (9) The general case is discussed e.g. in [51]; see also, for a more general discussion on normal forms with non-normal and non-semisimple normal form, the article ► [Perturbation of Systems with Nilpotent Real Part](#).
- (10) Note that the terms in  $\text{Ker}(\mathcal{A}) = \text{Ker}(\mathcal{A}^+)$  are just the resonant ones. Thus, the present characterization of normal forms is equivalent to the one given earlier on.
- (11) We also note that in the framework of Lie–Poincaré transformations, i.e. when considering the time one action of a vector field  $\Phi$  [25,28,145] (see above), Theorem 4 shows that  $\Phi$  can be chosen to admit  $Y$  as a symmetry; see [51].
- (12) Generalizations can also be obtained in the direction of relaxing the normality assumption for the matrices identifying the linear part of vector fields. We will not discuss this case, referring the reader instead to ► [Perturbation of Systems with Nilpotent Real Part](#), [51].
- (13) If this is the symmetry algebra of one of the vector fields, say  $X_0$ , we have that  $[X_0, X_i] = 0$  for all the  $i$ , i.e.  $X_0$  belongs to the center of the algebra  $\mathcal{G}$ .
- (14) We stress that we refer here to the case of a nilpotent Lie algebra, not to the case where the relevant matrices are nilpotent!
- (15) This is readily shown by the following example. Consider the two-dimensional system, which is *not* in normal form,  $\dot{x}_1 = -x_1 x_2$ ,  $\dot{x}_2 = x_2$ . It is easily seen that  $\mu = x_1 \exp(x_2)$  is a constant of motion, but not of the linearized problem, i.e.  $\mu \in \mathcal{I}_X$ , but  $\mu \notin \mathcal{I}_A$ . Similarly, we have that  $Y = x_1^2 \exp(x_2) (\partial / \partial x_1) \in \mathcal{G}_X$  but  $Y \notin \mathcal{G}_A$ .
- (16) It should be noted that  $X_k$ , and hence  $\mathcal{L}_k$ , change under further normalization transformations; however, they stabilize after a finite number of steps, and in particular will not change anymore after the further normalization reaches their order.
- (17) Note that the  $\widetilde{X}_k$  in Theorem 12 satisfies  $\mathcal{B}(\widetilde{X}_k) := [Y, \widetilde{X}_k] = [Y, X_k] + [Y, [X_0, Q_k]]$ ; using the Jacobi identity, and assuming  $[Y, Q_k] = 0$ , this reads  $\mathcal{B}(X_k) - \mathcal{A}[\mathcal{B}(Q_k)]$ . On the other hand,  $[A, B] = 0$  guarantees  $[\mathcal{A}, \mathcal{B}] = 0$ , hence we get  $\mathcal{B}(\widetilde{X}_k) = \mathcal{B}[X_k - \mathcal{A}(Q_k)]$ . Thus the assumption of Theorem 12 could be rephrased in terms of kernels of the operators  $\mathcal{A}^+$  and  $\mathcal{B}$  as follows: for each  $X_k \in \text{Ker}(\mathcal{A}^+) \cap \mathcal{V}_k$ , there exists  $Q_k \in \text{Ker}(\mathcal{B})$  such that  $[X_k - \mathcal{A}(Q_k)] \in \text{Ker}(\mathcal{B})$ .

- (18) As the  $\varphi$  identify group orbits for the group  $G$  generated by the Lie algebra, we interpret  $\dot{\varphi} = z(\varphi)$  as an equation in orbit space, and the equation for  $(x, w)$  as an equation on the Lie group  $G$ . Methods for the solution of the latter are discussed in [190], see also [38].

## Bibliography

1. Abenda S, Gaeta G, Walcher S (eds) (2003) Symmetry and Perturbation Theory – SPT2002. In: Proceedings of Cala Gonone workshop, 19–26 May 2002. World Scientific, Singapore
2. Abud M, Sartori G (1983) The geometry of spontaneous symmetry breaking. *Ann Phys* 150:307–372
3. Alekseevskij DV, Vinogradov AM, Lychagin VV (1991) Basic ideas and concepts of differential geometry. In: Gamkrelidze RV (ed) *Encyclopaedia of Mathematical Sciences vol 28 – Geometry I*. Springer, Berlin
4. Arnal D, Ben Ammar M, Pinczon G (1984) The Poincaré–Dulac theorem for nonlinear representations of nilpotent Lie algebras. *Lett Math Phys* 8:467–476
5. Arnold VI (1974) *Equations différentielles ordinaires*. MIR, Moscow, 2nd edn 1990. Arnold VI (1992) *Ordinary Differential Equations*. Springer, Berlin
6. Arnold V (1976) *Les méthodes mathématiques de la mécanique classique*. MIR, Moscow. Arnold VI (1983, 1989) *Mathematical methods of classical Mechanics*. Springer, Berlin
7. Arnold V (1980) *Chapitres supplémentaires de la théorie des équations différentielles ordinaires*. MIR, Moscow. Arnold VI (1983) *Geometrical methods in the theory of ordinary differential equations*. Springer, Berlin
8. Arnold VI, Il'yashenko YS (1988) Ordinary differential equations. In: Anosov DV, Arnold VI (eds) *Encyclopaedia of Mathematical Sciences vol 1 – Dynamical Systems I*, pp 1–148. Springer, Berlin
9. Arnold VI, Kozlov VV, Neishtadt AI (1993) Mathematical aspects of classical and celestial mechanics. In: Arnold VI (ed) *Encyclopaedia of Mathematical Sciences vol 3 – Dynamical Systems III*, 2nd edn, pp 1–291. Springer, Berlin
10. Baider A (1989) Unique normal form for vector fields and Hamiltonians. *J Diff Eqs* 78:33–52
11. Baider A, Churchill RC (1988) Uniqueness and non-uniqueness of normal forms for vector fields. *Proc R Soc Edinburgh A* 108:27–33
12. Baider A, Sanders J (1992) Further reduction of the Takens–Bogdanov normal form. *J Diff Eqs* 99:205–244
13. Bakri T, Nabergoj R, Tondl A, Verhulst F (2004) Parametric excitation in non-linear dynamics. *Int J Nonlin Mech* 39:311–329
14. Bambusi D, Gaeta G (eds) (1997) *Symmetry and Perturbation Theory*. In: Proceedings of Torino Workshop, ISI, December 1996. GNFM–CNR, Roma
15. Bambusi D, Gaeta G (2002) On persistence of invariant tori and a theorem by Nekhoroshev. *Math Phys El J* 8:1–13
16. Bambusi D, Cicogna G, Gaeta G, Marmo G (1998) Normal forms, symmetry, and linearization of dynamical systems. *J Phys A Math Gen* 31:5065–5082
17. Bambusi D, Gaeta G, Cadoni M (eds) (2001) *Symmetry and Perturbation Theory – SPT2001*. In: Proceedings of the international conference SPT2001, Cala Gonone, 6–13 May 2001. World Scientific, Singapore
18. Bargmann V (1961) On a Hilbert space of analytic functions and an associated integral transform. *Comm Pure Appl Math* 14:187–214
19. Baumann G (2000) *Symmetry analysis of differential equations with Mathematica*. Springer, New York
20. Belitskii GR (1978) Equivalence and normal forms of germs of smooth mappings. *Russ Math Surveys* 33(1):107–177
21. Belitskii GR (1981) Normal forms relative to the filtering action of a group. *Trans Moscow Math Soc* 40(2):1–39
22. Belitskii GR (1987) Smooth equivalence of germs of vector fields with a single eigenvalue or a pair of purely imaginary eigenvalues. *Funct Anal Appl* 20:253–259
23. Belitskii GR (2002)  $C^\infty$ -Normal forms of local vector fields. *Acta Appl Math* 70:23–41
24. Belmonte C, Boccaletti D, Pucacco G (2006) Stability of axial orbits in galactic potentials. *Cel Mech Dyn Astr* 95:101–116
25. Benettin G, Galgani L, Giorgilli A (1984) A proof of the Kolmogorov theorem on invariant tori using canonical transformations defined by the Lie method. *Nuovo Cimento B* 79:201–223
26. Bluman GW, Anco SC (2002) *Symmetry and integration methods for differential equations*. Springer, Berlin
27. Bluman GW, Kumei S (1989) *Symmetries and differential equations*. Springer, Berlin
28. Bogoliubov NN, Mitropolsky VA (1961) *Asymptotic methods in the theory of nonlinear oscillations*. Hindustan, New Delhi. (1962) *Méthodes asymptotiques dans la théorie des oscillations non-linéaires*. Gauthier-Villars, Paris
29. Broer HW (1979) Bifurcations of singularities in volume preserving vector fields. Ph D Thesis, Groningen
30. Broer HW (1981) Formal normal form theorems for vector fields and some consequences for bifurcations in the volume preserving case. In: Rand DA, Young LS (eds) *Dynamical systems and turbulence*. *Lect Notes Math* 898. Springer, Berlin
31. Broer HW, Takens F (1989) Formally symmetric normal forms and genericity. *Dyn Rep* 2:39–59
32. Bryuno AD (1971) Analytical form of differential equations I. *Trans Moscow Math Soc* 25:131–288
33. Bryuno AD (1971) Analytical form of differential equations II. *Trans Moscow Math Soc* 26:199–239
34. Bryuno AD (1988) The normal form of a Hamiltonian system. *Russ Math Sur* 43(1):25–66
35. Bryuno AD (1989) *Local Methods in the Theory of Differential Equations*. Springer, Berlin
36. Bryuno AD, Walcher S (1994) Symmetries and convergence of normalizing transformations. *J Math Anal Appl* 183:571–576
37. Cantwell BJ (2002) *Introduction to Symmetry Analysis*. Cambridge University Press, Cambridge
38. Carinena JF, Grabowski J, Marmo G (2000) *Lie-Scheffers systems: a geometric approach*. Bibliopolis, Napoli
39. Chen G, Della Dora J (2000) Further reductions of normal forms for dynamical systems. *J Diff Eqs* 166:79–106
40. Chern SS, Chen WH, Lam KS (1999) *Lectures on differential geometry*. World Scientific, Singapore
41. Chossat P (2002) The reduction of equivariant dynamics to the orbit space for compact group actions. *Acta Appl Math* 70:71–94
42. Chossat P, Lauterbach R (1999) *Methods in equivariant bifurcations and dynamical systems with applications*. World Scientific, Singapore



43. Chow SN, Hale JK (1982) *Methods of bifurcation theory*. Springer, Berlin
44. Chow SN, Li C, Wang D (1994) *Normal forms and bifurcations of planar vector fields*. Cambridge University Press, Cambridge
45. Chua LO, Kokubu H (1988) Normal forms for nonlinear vector fields Part I: theory. *IEEE Trans Circ Syst* 35:863–880
46. Chua LO, Kokubu H (1989) Normal forms for nonlinear vector fields Part II: applications. *IEEE Trans Circ Syst* 36:851–870
47. Churchill RC, Kummer M, Rod DL (1983) On averaging, reduction and symmetry in Hamiltonian systems. *J Diff Eqs* 49:359–414
48. Cicogna G, Gaeta G (1994) Normal forms and nonlinear symmetries. *J Phys A* 27:7115–7124
49. Cicogna G, Gaeta G (1994) Poincaré normal forms and Lie point symmetries. *J Phys A* 27:461–476
50. Cicogna G, Gaeta G (1994) Symmetry invariance and center manifolds in dynamical systems. *Nuovo Cim B* 109:59–76
51. Cicogna G, Gaeta G (1999) *Symmetry and perturbation theory in nonlinear dynamics*. Springer, Berlin
52. Cicogna G, Walcher S (2002) Convergence of normal form transformations: the role of symmetries. *Acta Appl Math* 70:95–111
53. Courant R, Hilbert D (1962) *Methods of Mathematical Physics*. Wiley, New York; (1989)
54. Cushman R, Sanders JA (1986) Nilpotent normal forms and representation theory of  $s(2, R)$ . In: Golubitsky M, Guckenheimer J (eds) *Multi-parameter bifurcation theory*. *Contemp Math* 56, AMS, Providence,
55. Crawford JD (1991) Introduction to bifurcation theory. *Rev Mod Phys* 63:991–1037
56. Crawford JD, Knobloch E (1991) Symmetry and symmetry-breaking bifurcations in fluid dynamics. *Ann Rev Fluid Mech* 23:341–387
57. Degasperis A, Gaeta G (eds) (1999) *Symmetry and Perturbation Theory II – SPT98*. In: *Proceedings of Roma Workshop, Università La Sapienza, December 1998*. World Scientific, Singapore
58. Deprit A (1969) Canonical transformation depending on a small parameter. *Celest Mech* 1:12–30
59. de Zeeuw T, Merritt D (1983) Stellar orbits in a triaxial galaxy I. Orbits in the plane of rotation. *Astrophys J* 267:571–595
60. Elphick C, Tirapegui E, Brachet ME, Coullet P, Iooss G (1987) A simple global characterization for normal forms of singular vector fields. *Physica D* 29:95–127. (1988) Addendum. *Physica D* 32:488
61. Fassò F (1990) Lie series method for vector fields and Hamiltonian perturbation theory. *ZAMP* 41:843–864
62. Fassò F, Guzzo M, Benettin G (1998) Nekhoroshev stability of elliptic equilibria of Hamiltonian systems. *Comm Math Phys* 197:347–360
63. Field MJ (1989) Equivariant bifurcation theory and symmetry breaking. *J Dyn Dif Eqs* 1:369–421
64. Field MJ (1996) *Lectures on bifurcations, dynamics and symmetry*. *Res Notes Math* 356. Pitman, Boston
65. Field MJ (1996) Symmetry breaking for compact Lie groups. *Mem AMS* 574:1–170
66. Field MJ, Richardson RW (1989) Symmetry breaking and the maximal isotropy subgroup conjecture for reflection groups. *Arch Rat Mech Anal* 105:61–94
67. Field MJ, Richardson RW (1990) Symmetry breaking in equivariant bifurcation problems. *Bull Am Math Soc* 22:79–84
68. Field MJ, Richardson RW (1992) Symmetry breaking and branching patterns in equivariant bifurcation theory I. *Arch Rat Mech Anal* 118:297–348
69. Field MJ, Richardson RW (1992) Symmetry breaking and branching patterns in equivariant bifurcation theory II. *Arch Rat Mech Anal* 120:147–190
70. Fokas AS (1979) Generalized symmetries and constants of motion of evolution equations. *Lett Math Phys* 3:467–473
71. Fokas AS (1979) Group theoretical aspects of constants of motion and separable solutions in classical mechanics. *J Math Anal Appl* 68:347–370
72. Fokas AS (1980) A symmetry approach to exactly solvable evolution equations. *J Math Phys* 21:1318–1326
73. Fokas AS (1987) Symmetries and integrability. *Stud Appl Math* 77:253–299
74. Fokas AS, Gelfand IM (1996) Surfaces on Lie groups, Lie algebras, and their integrability. *Comm Math Phys* 177:203–220
75. Fontich E, Gelfreich VG (1997) On analytical properties of normal forms. *Nonlinearity* 10:467–477
76. Forest E, Murray D (1994) Freedom in minimal normal forms. *Physica D* 74:181–196
77. Fushchich WI, Nikitin AG (1987) *Symmetries of Maxwell equations*. Reidel, Dordrecht
78. Fushchich WI, Shtelen WM, Slavutsky SL (1989) *Symmetry analysis and exact solutions of nonlinear equations of mathematical physics*. Naukova Dumka, Kiev
79. Gaeta G (1990) Bifurcation and symmetry breaking. *Phys Rep* 189:1–87
80. Gaeta G (1994) *Nonlinear symmetries and nonlinear equations*. Kluwer, Dordrecht
81. Gaeta G (1997) Reduction of Poincaré normal forms. *Lett Math Phys* 42:103–114 & 235
82. Gaeta G (1999) An equivariant branching lemma for relative equilibria. *Nuovo Cim B* 114:973–982
83. Gaeta G (1999) Poincaré renormalized forms. *Ann IHP Phys Theor* 70:461–514
84. Gaeta G (2001) Algorithmic reduction of Poincaré-Dulac normal forms and Lie algebraic structure. *Lett Math Phys* 57:41–60
85. Gaeta G (2002) Poincaré normal and renormalized forms. *Acta Appl Math* 70:113–131
86. Gaeta G (2002) Poincaré normal forms and simple compact Lie groups. *Int J Mod Phys A* 17:3571–3587
87. Gaeta G (2002) The Poincaré–Lyapounov–Nekhoroshev theorem. *Ann Phys* 297:157–173
88. Gaeta G (2003) The Poincaré–Nekhoroshev map. *J Nonlin Math Phys* 10:51–64
89. Gaeta G (2006) Finite group symmetry breaking. In: Francoise JP, Naber G, Tsou ST (eds) *Encyclopedia of Mathematical Physics*. Kluwer, Dordrecht
90. Gaeta G (2006) Non-quadratic additional conserved quantities in Birkhoff normal forms. *Cel Mech Dyn Astr* 96:63–81
91. Gaeta G (2006) The Poincaré–Lyapounov–Nekhoroshev theorem for involutory systems of vector fields. *Ann Phys NY* 321:1277–1295
92. Gaeta G, Marmo G (1996) Nonperturbative linearization of dynamical systems. *J Phys A* 29:5035–5048
93. Gaeta G, Morando P (1997) Michel theory of symmetry breaking and gauge theories. *Ann Phys NY* 260:149–170



94. Gaeta G, Walcher S (2005) Dimension increase and splitting for Poincaré-Dulac normal forms. *J Nonlin Math Phys* 12:S1327-S1342
95. Gaeta G, Walcher S (2006) Embedding and splitting ordinary differential equations in normal form. *J Diff Eqs* 224:98-119
96. Gaeta G, Prinari B, Rauch S, Terracini S (eds) (2005) *Symmetry and Perturbation Theory – SPT2004*. In: Proceedings of Cala Gonone workshop, 30 May – 6 June 2004. World Scientific, Singapore
97. Gaeta G, Grosshans FD, Scheurle J, Walcher S (2008) Reduction and reconstruction for symmetric ordinary differential equations. *J Diff Eqs* 244:1810-1839
98. Gaeta G, Vitolo R, Walcher S (eds) (2007) *Symmetry and Perturbation Theory – SPT2007*. In: Proceedings of Otranto workshop, 2-9 June 2007. World Scientific, Singapore
99. Gallavotti G (1983) *The elements of mechanics*. Springer, Berlin
100. Giorgilli A (1988) Rigorous results on the power expansions for the integrals of a Hamiltonian system near an elliptic equilibrium point. *Ann IHP Phys Theor* 48:423-439
101. Giorgilli A, Locatelli U (1997) Kolmogorov theorem and classical perturbation theory. *ZAMP* 48:220-261
102. Giorgilli A, Morbidelli A (1997) Invariant KAM tori and global stability for Hamiltonian systems. *ZAMP* 48:102-134
103. Giorgilli A, Zehnder E (1992) Exponential stability for time dependent potentials. *ZAMP* 43:827-855
104. Glendinning P (1994) *Stability, instability and chaos: an introduction to the theory of nonlinear differential equations*. Cambridge University Press, Cambridge
105. Golubitsky M, Stewart I, Schaeffer D (1988) *Singularity and groups in bifurcation theory – vol II*. Springer, Berlin
106. Gramchev T, Yoshino M (1999) Rapidly convergent iteration methods for simultaneous normal forms of commuting maps. *Math Z* 231:745-770
107. Guckenheimer J, Holmes P (1983) *Nonlinear oscillations, dynamical systems, and bifurcation of vector fields*. Springer, Berlin
108. Gustavson FG (1964) On constructing formal integrals of a Hamiltonian system near an equilibrium point *Astron J* 71:670-686
109. Guzzo M, Fassò F, Benettin G (1998) On the stability of elliptic equilibria. *Math Phys EI J* 4(1):16
110. Hamermesh M (1962) *Group theory*. Addison-Wesley, Reading; reprinted by Dover, New York (1991)
111. Hanssmann H (2007) *Local and semi-local bifurcations in Hamiltonian dynamical systems Results and examples*. Springer, Berlin
112. Hermann R (1968) The formal linearization of a semisimple Lie algebra of vector fields about a singular point. *Trans AMS* 130:105-109
113. Hoveijn I (1996) Versal deformations and normal forms for reversible and Hamiltonian linear systems. *J Diff Eq* 126:408-442
114. Hoveijn I, Verhulst F (1990) Chaos in the 1:2:3 Hamiltonian normal form. *Physica D* 44:397-406
115. Hydor PE (2000) *Symmetry methods for differential equations*. Cambridge UP, Cambridge
116. Ibragimov N (1992) Group analysis of ordinary differential equations and the invariance principle in Mathematical Physics. *Russ Math Surv* 47(4):89-156
117. Il'yashenko YS, Yakovenko SY (1991) *Finitely smooth normal forms of local families of diffeomorphisms and vector fields*. *Russ Math Surv* 46(1):1-43
118. Iooss G, Adelmeyer M (1992) *Topics in bifurcation theory and applications*. World Scientific, Singapore
119. Isham CJ (1999) *Modern differential geometry for physicists*. World Scientific, Singapore
120. Kinyon M, Walcher S (1997) On ordinary differential equations admitting a finite linear group of symmetries. *J Math Analysis Appl* 216:180-196
121. Kirillov AA (1976, 1984) *Elements of the Theory of Representations*. Springer, Berlin
122. Kodama Y (1994) Normal forms, symmetry and infinite dimensional Lie algebra for systems of ODE's. *Phys Lett A* 191:223-228
123. Kokubu H, Oka H, Wang D (1996) Linear grading function and further reduction of normal forms. *J Diff Eq* 132:293-318
124. Krasil'shchik IS, Vinogradov AM (1984) Nonlocal symmetries and the theory of coverings. *Acta Appl Math* 2:79-96
125. Krasil'shchik IS, Vinogradov AM (1999) *Symmetries and conservation laws for differential equations of mathematical physics*. AMS, Providence
126. Kummer M (1971) How to avoid secular terms in classical and quantum mechanics. *Nuovo Cimento B* 1:123-148
127. Kummer M (1976) On resonant nonlinearly coupled oscillators with two equal frequencies. *Comm Math Phys* 48:53-79
128. Lamb J (1996) Local bifurcations in  $k$ -symmetric dynamical systems. *Nonlinearity* 9:537-557
129. Lamb J (1998)  $k$ -symmetry and return maps of spacetime symmetric flows. *Nonlinearity* 11:601-630
130. Lamb J, Melbourne I (2007) Normal form theory for relative equilibria and relative periodic solutions. *Trans AMS* 359:4537-4556
131. Lamb J, Roberts J (1998) Time reversal symmetry in dynamical systems: a survey. *Physica D* 112:1-39
132. Levi D, Winternitz P (1989) Non-classical symmetry reduction: example of the Boussinesq equation. *J Phys A* 22:2915-2924
133. Lin CM, Vittal V, Kliemann W, Fouad AA (1996) Investigation of modal interaction and its effect on control performance in stressed power systems using normal forms of vector fields. *IEEE Trans Power Syst* 11:781-787
134. Marsden JE (1992) *Lectures on Mechanics*. Cambridge University Press, Cambridge
135. Marsden JE, Ratiu T (1994) *Introduction to mechanics and symmetry*. Springer, Berlin
136. Michel L (1971) Points critiques de fonctions invariantes sur une  $G$ -variété. *Comptes Rendus Acad Sci Paris* 272-A:433-436
137. Michel L (1971) Nonlinear group action Smooth action of compact Lie groups on manifolds. In: Sen RN, Weil C (eds) *Statistical Mechanics and Field Theory*. Israel University Press, Jerusalem
138. Michel L (1975) Les brisure spontanées de symétrie en physique. *J Phys Paris* 36-C7:41-51
139. Michel L (1980) Symmetry defects and broken symmetry Configurations Hidden symmetry. *Rev Mod Phys* 52:617-651
140. Michel L, Radicati L (1971) Properties of the breaking of hadronic internal symmetry. *Ann Phys NY* 66:758-783
141. Michel L, Radicati L (1973) The geometry of the octet. *Ann IHP* 18:185-214
142. Michel L, Zhilinskii BI (2001) Symmetry, invariants, topology Basic tools. *Phys Rep* 341:11-84
143. Mikhailov AV, Shabat AB, Yamilov RI (1987) *The symmetry ap-*

- proach to the classification of non-linear equations Complete list of integrable systems. *Russ Math Surv* 42(4):1–63
144. Meyer KR, Hall GR (1992) Introduction to Hamiltonian dynamical systems and the N-body problem. Springer, New York
  145. Mitropolsky YA, Lopatin AK (1995) Nonlinear mechanics, groups and symmetry. Kluwer, Dordrecht
  146. Nakahara M (1990) Geometry, Topology and Physics. IOP, Bristol
  147. Nash C, Sen S (1983) Topology and geometry for physicists. Academic Press, London
  148. Nekhoroshev NN (1994) The Poincaré–Lyapunov–Liouville–Arnol'd theorem. *Funct Anal Appl* 28:128–129
  149. Nekhoroshev NN (2002) Generalizations of Gordon theorem. *Regul Chaotic Dyn* 7:239–247
  150. Nekhoroshev NN (2005) Types of integrability on a submanifold and generalizations of Gordons theorem. *Trans Moscow Math Soc* 66:169–241
  151. Olver PJ (1986) Applications of Lie groups to differential equations. Springer, Berlin
  152. Olver PJ (1995) Equivalence, Invariants, and Symmetry. Cambridge University Press, Cambridge
  153. Ovsjiannikov LV (1982) Group analysis of differential equations. Academic Press, London
  154. Palacián J, Yanguas P (2000) Reduction of polynomial Hamiltonians by the construction of formal integrals. *Nonlinearity* 13:1021–1054
  155. Palacián J, Yanguas P (2001) Generalized normal forms for polynomial vector fields. *J Math Pures Appl* 80:445–469
  156. Palacián J, Yanguas P (2003) Equivariant N-DOF Hamiltonians via generalized normal forms. *Comm Cont Math* 5:449–480
  157. Palacián J, Yanguas P (2005) A universal procedure for normalizing  $n$ -degree-of-freedom polynomial Hamiltonian systems. *SIAM J Appl Math* 65:1130–1152
  158. Pucci E, Saccomandi G (1992) On the weak symmetry group of partial differential equations. *J Math Anal Appl* 163:588–598
  159. Ruelle D (1973) Bifurcation in the presence of a symmetry group. *Arch Rat Mech Anal* 51:136–152
  160. Ruelle D (1989) Elements of Differentiable Dynamics and Bifurcation Theory. Academic Press, London
  161. Sadovskii DA, Delos JB (1996) Bifurcation of the periodic orbits of Hamiltonian systems – an analysis using normal form theory. *Phys Rev A* 54:2033–2070
  162. Sanders JA (2003) Normal form theory and spectral sequences. *J Diff Eqs* 192:536–552
  163. Sanders JA (2005) Normal forms in filtered Lie algebra representations. *Acta Appl Math* 87:165–189
  164. Sanders JA, Verhulst F (1985) Averaging methods in nonlinear dynamical systems. Springer, Berlin
  165. Sanders JA, Verhulst F, Murdock J (2007) Averaging methods in nonlinear dynamical systems. Springer, Berlin
  166. Sartori G (1991) Geometric invariant theory A model-independent approach to spontaneous symmetry and/or supersymmetry breaking. *Riv N Cim* 14–11:1–120
  167. Sartori G (2002) Geometric invariant theory in a model-independent analysis of spontaneous symmetry and supersymmetry breaking. *Acta Appl Math* 70:183–207
  168. Sartori G, Valente G (2005) Constructive axiomatic approach to the determination of the orbit spaces of coregular compact linear groups. *Acta Appl Math* 87:191–228
  169. Sattinger DH (1979) Group theoretic methods in bifurcation theory. *Lecture Notes in Mathematics* 762. Springer, Berlin
  170. Sattinger DH (1983) Branching in the presence of symmetry. SIAM, Philadelphia
  171. Sattinger DH, Weaver O (1986) Lie groups and algebras. Springer, Berlin
  172. Siegel K, Moser JK (1971) Lectures on Celestial Mechanics. Springer, Berlin; reprinted in *Classics in Mathematics*. Springer, Berlin (1995)
  173. Sokolov VV (1988) On the symmetries of evolutions equations. *Russ Math Surv* 43(5):165–204
  174. Stephani H (1989) Differential equations Their solution using symmetries. Cambridge University Press, Cambridge
  175. Stewart I (1988) Bifurcation with symmetry. In: Bedford T, Swift J (eds) New directions in dynamical systems. Cambridge University Press, Cambridge
  176. Tondl A, Ruijgrok T, Verhulst F, Nabergoj R (2000) Autoparametric resonance in mechanical systems. Cambridge University Press, Cambridge
  177. Ushiki S (1984) Normal forms for singularities of vector fields. *Jpn J Appl Math* 1:1–34
  178. Vanderbauwhede A (1982) Local bifurcation and symmetry. Pitman, Boston
  179. Verhulst F (1989) Nonlinear differential equations and dynamical systems. Springer, Berlin; (1996)
  180. Verhulst F (1998) Symmetry and integrability in Hamiltonian normal form. In: Bambusi D, Gaeta G (eds) Symmetry and perturbation theory. CNR, Roma
  181. Verhulst F (1999) On averaging methods for partial differential equations. In: Degasperis A, Gaeta G (eds) Symmetry and perturbation theory II. World Scientific, Singapore
  182. Vinogradov AM (1984) Local symmetries and conservation laws. *Acta Appl Math* 2:21–78
  183. Vittal V, Kliemann W, Ni YX, Chapman DG, Silk AD, Sobajic DJ (1998) Determination of generator groupings for an islanding scheme in the Manitoba hydro system using the method of normal forms. *IEEE Trans Power Syst* 13:1346–1351
  184. Vorob'ev EM (1986) Partial symmetries of systems of differential equations. *Soviet Math Dokl* 33:408–411
  185. Vorob'ev EM (1991) Reduction and quotient equations for differential equations with symmetries. *Acta Appl Math* 23:1–24
  186. Walcher S (1991) On differential equations in normal form. *Math Ann* 291:293–314
  187. Walcher S (1993) On transformation into normal form. *J Math Anal Appl* 180:617–632
  188. Walcher S (1999) Orbital symmetries of first order ODEs. In: Degasperis A, Gaeta G (eds) Symmetry and perturbation theory II. World Scientific, Singapore
  189. Walcher S (2000) On convergent normal form transformations in the presence of symmetry. *J Math Anal Appl* 244:17–26
  190. Wei J, Norman E (1963) Lie algebraic solution of linear differential equations. *J Math Phys* 4:575–581
  191. Winternitz P (1987) What is new in the study of differential equations by group theoretical methods? In: Gilmore R (ed) Group Theoretical Methods in Physics proceedings of the XV ICGTMP. World Scientific, Singapore
  192. Winternitz P (1993) Lie groups and solutions of nonlinear PDEs. In: Ibort LA, Rodriguez MA (eds) Integrable systems, quantum groups, and quantum field theory NATO ASI 9009. Kluwer, Dordrecht

## Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence

TOM H. SOLOMON

Department of Physics & Astronomy,  
Bucknell University, Lewisburg, USA

### Article Outline

Glossary

Definition of the Subject

Introduction and Theoretical Background

Specific Examples of Patterns in Fluid Flows

Transport and Mixing

Pattern Formation in Reaction-Diffusion  
and Advection-Reaction-Diffusion Systems

Other Examples of Pattern Forming Systems

Future Directions

Bibliography

### Glossary

**Bifurcation** A change from one kind of behavior to another; e. g., a bifurcation from steady flow to time periodic flow.

**Defect** A location where a pattern abruptly changes; e. g., where a line ends or two lines meet at a point.

**Deterministic chaos** Erratic behavior arising from a low-dimensional, deterministic system; behavior that shows sensitive dependence on initial conditions where a small error in initial conditions grows roughly exponentially in time.

**Free energy functional** An effective energy that depends on the function describing a pattern.

**Fully-developed turbulence** Turbulent flow for very large Reynolds number with no spatial or temporal order.

**Grain boundary** Line or curve in a pattern where two different zones meet.

**Kolmogorov–Arnold–Moser (KAM) invariant surface** Curve or surface that separates region of ordered trajectories from region of chaotic trajectories. Tracers cannot cross between these two regions; consequently, a KAM invariant surface acts as a transport barrier.

**Laminar flow** Smooth, well-ordered flow, achieved with small Reynolds number.

**Lévy flights** Trajectories characterized by jumps between distant parts of the system. If it is a Lévy flight, the jumps in a trajectory have a probability distribution  $P(L) \sim L^{-\mu}$ , with  $2 < \mu < 3$ .

**Linear stability theory** Approach used to determine critical parameters (e. g., Re or Ra) at which a flow becomes unstable. Linear stability theory also determines the wavenumbers of the disturbances that are unstable.

**Poincaré section** A stroboscopic plot for a time-periodic flow showing the positions of one or more tracers once each period of the flow.

**Rayleigh–Bénard convection** A flow that develops from an instability due to unstable stratification when a fluid is heated from below.

**Rayleigh number** Dimensionless parameter that characterizes the magnitude of the temperature difference in a Rayleigh–Bénard flow. The Rayleigh number can also be thought of as a ratio of thermal buoyancy to diffusive damping.

**Reaction-diffusion system** A spatially-extended system that produces patterns or fronts as a result of the interaction between some sort of reaction (e. g., chemical reaction, biological interaction, phase transition) and molecular diffusion.

**Reynolds number** Dimensionless parameter that characterizes the strength of a flow. The Reynolds number can also be thought of as the ratio between fluid inertia and viscous damping.

**Stratification** Density variations in a fluid system.

**Strong turbulence** High Reynolds number turbulence.

**Subdiffusion** Transport where the variance  $\sigma^2$  grows more slowly than normal diffusion, e. g.,  $\sigma^2(t) \sim t^\gamma$ , with  $\gamma < 1$ . Associated with regions in the flow where the tracers temporarily stick.

**Superdiffusion** Transport where the variance  $\sigma^2$  grows more rapidly than normal diffusion, e. g.,  $\sigma^2(t) \sim t^\gamma$ , with  $\gamma > 1$ . Associated with Lévy flight trajectories.

**Taylor–Couette flow** Flow between two coaxial cylinders, the inner of which (at least) is rotating.

**Weak (defect-mediated) turbulence** Flows with intermediate values of the Reynolds number that still produce short-range patterns but evolve both spatially and temporally in a complicated manner. Can be characterized by the presence of numerous defects in the pattern (more defects for more turbulent flows) that move around the system.

### Definition of the Subject

“Nonlinear dynamics” is the study of systems described by equations of motion that depend nonlinearly on their variables. This is of significant scientific and mathematical interest because nonlinear systems – unlike simpler

linear ones – typically show quite complicated behavior, including deterministic chaos, spontaneous formation of patterns, spatio-temporal complexity and turbulence. The equations describing fluid flows are nonlinear; consequently, there is a wide variety of types of flows that can be achieved, spanning a wide range of levels of complexity from simple, well-ordered, *laminar* flows up through turbulent flows with significant complexity both in space and time. The equations that describe the mixing of impurities in a flow are also often nonlinear (depending on the fluid flow); consequently, mixing can also be very complicated and even *chaotic*.

### Introduction and Theoretical Background

The dynamics of a fluid flow are governed by the Navier–Stokes equation:

$$\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \vec{\nabla} \vec{u} = -\frac{1}{\rho} \vec{\nabla} p + \nu \nabla^2 \vec{u} + \frac{1}{\rho} \vec{F}.$$

This is the fluid continuum version of Newton’s second law. The first term on the left represents the change in momentum of a small parcel of a fluid. Several factors can result in a change in momentum: forces acting on the fluid element – pressure differences, viscosity and external forces (first, second and third terms on the right), and advection of higher or lower velocity fluid elements into the region of interest (the second term on the left). The Navier–Stokes equation can be non-dimensionalized by scaling the velocity by a characteristic velocity scale  $U$ , lengths by a typical length scale  $L$ , and time by an advective time scale  $L/U$ :

$$\frac{\partial \vec{u}'}{\partial t} + \vec{u}' \cdot \vec{\nabla}' \vec{u}' = -\vec{\nabla}'(\Delta p)' + \frac{1}{\text{Re}} \nabla'^2 \vec{u}',$$

where we have neglected external forcing and where  $\vec{u}' \equiv u/U$  is the non-dimensional velocity. The non-dimensional parameter  $\text{Re} = UL/\nu$  is the Reynolds number. The Reynolds number characterizes the ratio of inertia to viscosity in a fluid flow.

The advection term on the left is often referred to as the *inertial* term, and is the source of nonlinearity in the Navier–Stokes equation. Nonlinearity goes hand-in-hand with complexity; consequently, the Reynolds number is the fundamental parameter used to describe the transition in fluid flows from simple, laminar (smooth), well-ordered flows to more and more complicated flows. If  $\text{Re} \ll 1$ , the inertial term is negligible and for steady flows, viscosity at every location is balanced by pressure gradients. In this limit, the Navier–Stokes equations can be rewritten (in di-

mensional form) as:

$$\frac{1}{\rho} \vec{\nabla} p = \nu \nabla^2 \vec{u}.$$

Flows in this limit are called *viscous*, *creeping* or *Stokes* flows. If  $\text{Re} \gg 1$ , viscosity is negligible and the NS Equation can be written as:

$$\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \vec{\nabla} \vec{u} = -\frac{1}{\rho} \vec{\nabla} p.$$

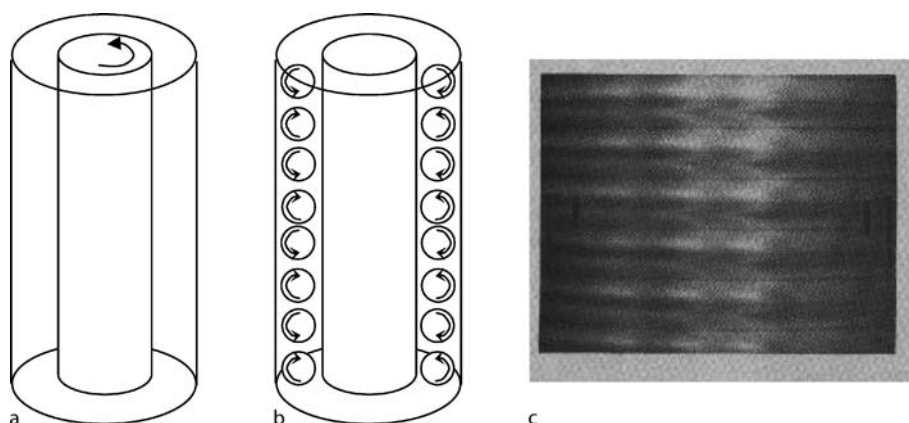
Flows in this limit are called *inviscid* or *Euler* flows and are dominated by inertial effects. Flows with very large  $\text{Re}$  are typically turbulent.

There has been significant research into the transition between well-ordered, laminar, viscous flows for  $\text{Re} \ll 1$  and fully-developed turbulent flows for  $\text{Re} \gg 1$ . Landau [1] predicted that an increase in  $\text{Re}$  would be accompanied by a infinite series of bifurcations as more and more frequencies of fluid oscillation (and corresponding spatial complexity) are added to the flow. According to Landau, measurement of the flow velocity at a point will reveal a steady, time-independent signal for weak flows (low  $\text{Re}$ ), then for stronger flows, the velocity will oscillate periodically at that point, then at higher  $\text{Re}$ , the velocity time series at that point will be a superposition of two periodic oscillations with different frequencies. As  $\text{Re}$  increases, the velocity time series will be a superposition of more and more sinusoidal functions with different frequencies. Ultimately, according to Landau, a turbulent flow is simply one with so many superposed frequencies that it *seems* complicated and unpredictable.

Landau’s prediction was tested in the early 1970s by Harry Swinney and Jerry Gollub in a *Taylor–Couette* flow [2] (see Subject. “[Taylor–Couette Flow](#)” for more detail about Taylor–Couette flow). However, they did not find an infinite sequence of transitions with more and more frequencies (and more and more complexity). Instead, they found a sudden transition from an ordered flow with two superposed frequencies directly to broad spectrum, aperiodic time dependence, a state that we now know to be *chaotic*. This direct transition in fluid flows to chaotic time dependence was also predicted theoretically by Ruelle and Takens in 1971 [3].

*Deterministic chaos* is denoted by two main features: (1) complicated, aperiodic time dependence from a simple, deterministic system with a small number of degrees of freedom; and (2) sensitive dependence on initial conditions, whereby small uncertainties in measurements of the initial conditions grow roughly exponentially in time, destroying long-term predictability. Despite the name, however, chaotic systems are usually quite well-ordered spa-





**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 1**

**Taylor–Couette flow.** **a** Diagram of configuration. The fluid is contained in the gap between two coaxial cylinders, the inner of which (at least) rotates, driving a flow between the two cylinders. **b** Cross-section sketch of Taylor–Couette vortices. **c** Experimental image showing Taylor–Couette vortices; flow is visualized using Kalliroscope. (From [6])

tially; it is the time-dependence that is complicated. Further increases in  $Re$  are needed to break up the spatial patterns that persist for chaotic flows. A regime displaying disorder in both space and time is referred to as *spatio-temporal chaos* or *weak turbulence*. Weak turbulent regimes are still characterized by well-defined structures (e.g., vortices), but the structures are arranged in a disordered pattern. These structures break up if  $Re$  is increased beyond that for the weak turbulent regime; ultimately, for large enough  $Re$ , the flow is described as *fully-developed turbulence* with a broad spectrum of spatial length scales. Long-lived vortex structures may survive for surprisingly large  $Re$ , however, especially in planetary and atmospheric flows in rotating systems [4] or in plasma flows with large uniform magnetic fields.

In this article, we present several different examples of pattern formation, mixing and chaos in fluid systems. Section “[Specific Examples of Patterns in Fluid Flows](#)” presents specific examples of the kinds of transitions discussed above in fluid flows, showing the patterns that accompany these transitions. We emphasize Taylor–Couette and convective flows, mainly due to their historical importance in this field. In Sect. “[Transport and Mixing](#)”, we discuss transport and mixing, highlighting in particular the relatively recent discovery that fluid mixing in *laminar*, well-ordered flows is often chaotic. In Sect. “[Pattern Formation in Reaction-Diffusion and Advection-Reaction-Diffusion Systems](#)”, we discuss chemical pattern formation in both stagnant and flowing systems. Section “[Other Examples of Pattern Forming Systems](#)” contains a brief listing of some other types of pattern-forming systems. We finish in Sect. “[Future Directions](#)” with a brief overview of future directions in the field.

### Specific Examples of Patterns in Fluid Flows

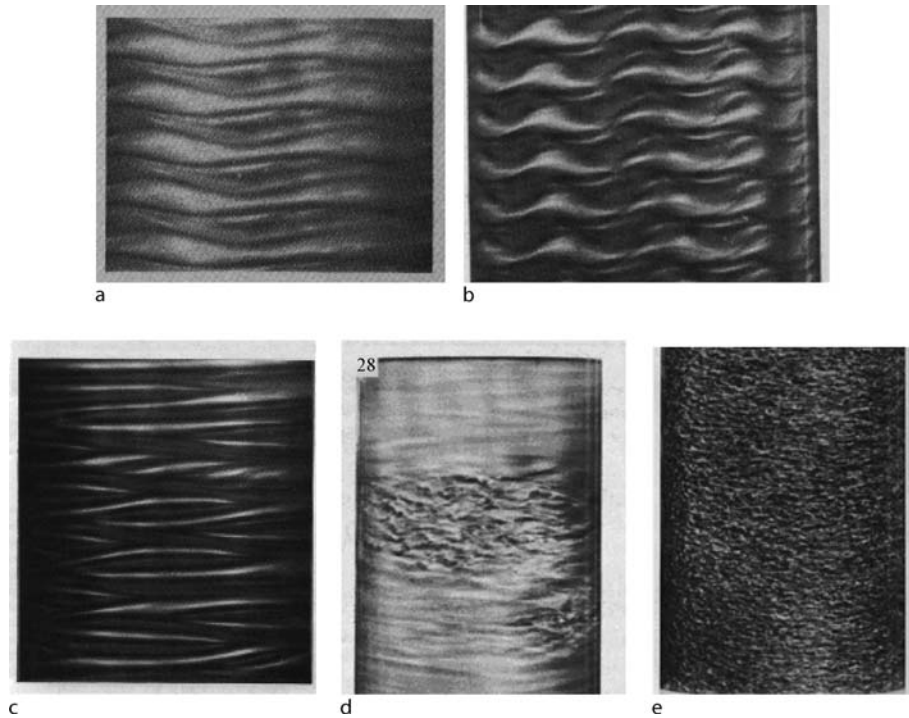
There are a number of different flows whose spatial patterns have been studied. Similar patterns are found in a very wide range of flows.

#### Taylor–Couette Flow

Taylor–Couette flow is the flow between two coaxial cylinders (Fig. 1a). In its simplest form, the outer cylinder remains stationary while the inner cylinder rotates. No-slip boundary conditions requires fluid at the outer edge to remain stationary while fluid at the inner edge rotates with the inner cylinder. The Reynolds number for this flow is defined using the gap width  $L$  as the length scale and the maximum flow velocity  $U$  as the velocity scale:  $Re = UL/\nu$ . Taylor–Couette flows are often visualized with Kalliroscope [5], a suspension of aluminum flakes that align with shear zones in the flow. Reflection of light off the flakes depends on their orientation; consequently, kalliroscope quite effectively visualizes the flow field. This technique was originally developed by the artist Paul Matisse, but has been adopted by experimentalists who study fluid instabilities.

For small  $Re$  in the Taylor–Couette system, flow between the cylinders is laminar and purely in the azimuthal direction, with the velocity magnitude dropping smoothly from the inner cylinder velocity to zero at the outer cylinder. Beyond a critical  $Re$  this smooth flow becomes unstable due to centrifugal effects – the faster moving fluid near the inner cylinder moves outward while the slower moving fluid near the outside moves inward. The result is a pattern of stacked tori (donut-shaped structures) called *Taylor vortices* [6] (Fig. 1b and c).





**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 2**

Some examples of Taylor–Couette patterns; all visualized with Kalliroscope flakes. **a** Wavy-vortex flow (from [6]); **b** Modulated wavy vortex flow (from [7]); **c** Braided vortex flow (from [8]); **d** Intermittent turbulent spots (from [7]); and **e** Turbulent flow (from [7])

A fluid element within a Taylor vortex still moves azimuthally around the cylinders, but also follows a closed (vortex) path in  $r$ – $z$  coordinates. A radial cross-section of the flow reveals a chain of counter-rotating vortices.

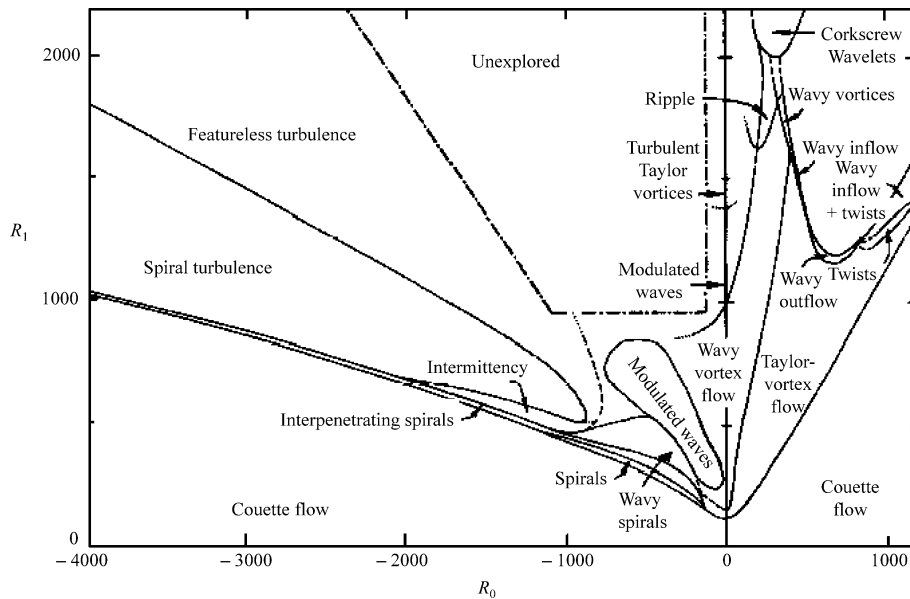
Further increases in  $Re$  result in a bifurcation to *wavy-vortex* flow [7], where the tori defining the Taylor vortices develop a snake-like undulation that propagates around the annulus (Fig. 2a). This spatial pattern corresponds to the time-periodic state discussed in transition to chaos in Sect. “[Introduction and Theoretical Background](#)”. Further increases in  $Re$  result in transitions to quasi-periodic and chaotic states. Ultimately, for large enough  $Re$ , the flow is turbulent (Fig. 2e).

If the outer cylinder rotates as well, a wide variety of spatial patterns are found [7,8]. Some of these states are shown in Fig. 2. Taylor–Couette flows have been studied extensively because of the wide variety of different states, including states in which the flow experiences intermittent (momentary) bursts of turbulence. Various investigators have mapped out *parameter space diagrams* that show the conditions needed to obtain various different states in Taylor–Couette flow. An example of such a parameter space diagram [7] is shown in Fig. 3.

### Rayleigh–Bénard Convection

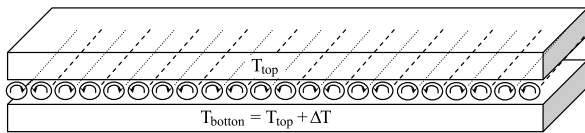
When a fluid is heated from below or from the side, the warmer fluid near the bottom expands and becomes less dense than the cooler fluid above. This results in unstable stratification with heavier, more dense fluid above lighter, less dense fluid. If the stratification is large enough, the denser fluid falls and the lighter fluid rises, producing a *convective* flow. Thermal convection of this nature is quite common: it occurs when heating up water or soup on a stove; it is a dominant mechanism by which a heating element heats up the air in an oven or by which heat from a radiator warms up a room; it is the dominant mechanism by which atmospheric flows are driven; it is the mechanism by which flows within the Earth’s crust are formed; and it is a major driving force for flows in stars. In all cases, thermal convection results in *significantly* faster and more efficient transport of heat than would be achieved solely with thermal diffusion.

In its simplest form, thermal convection ensues if a large enough vertical temperature gradient is applied with the bottom warmer than the top (Fig. 4). The relevant dimensionless number characterizing the magnitude of the temperature difference is the *Rayleigh Number*



Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 3

Parameter-space diagram showing the many different regimes of pattern formation for Taylor–Couette flow between two coaxial cylinders. (From [7])



Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 4

Schematic for Rayleigh–Bénard convection. A thin layer of fluid is sandwiched between two thermally-conducting plates, with the bottom plate hotter than the top plate. For sufficiently large  $\Delta T$ , convection ensues with hotter, light fluid rising from the bottom and colder, denser fluid falling from the top. In its simplest form, the convection pattern is a pattern of counter-rotating "rolls" as shown in the sketch. The dotted (dashed) lines denote "downwelling" (upwelling) regions in the flow

$Ra = g\alpha\Delta T/\nu\kappa$ , where  $\Delta T = T_{\text{bottom}} - T_{\text{top}}$  is the temperature gradient,  $\alpha$  is the coefficient describing the volume expansion of the fluid in response to a change in temperature,  $g$  is the gravitational acceleration and  $\kappa$  is the thermal diffusivity. Conceptually, increases in the temperature difference, gravitational acceleration, or expansion coefficient make the fluid more unstable, whereas viscosity inhibits fluid motion, and thermal diffusivity tends to diminish temperature differences.

If  $Ra$  is small, the fluid remains motionless – viscosity and thermal diffusion are sufficient to keep the fluid from becoming unstable. If  $Ra$  exceed a critical value  $Ra_c$ , the motionless state becomes unstable and the warmer, less-

dense fluid near the bottom rises while the cooler, more-dense fluid near the top sinks. In a thin, rectangular box a pattern of *convection rolls* develops, with fluid circulating like rotating cigars in a box, as shown schematically in Fig. 4.

Numerous studies have been made of patterns made by Rayleigh–Benard convection. In a manner similar to that in Taylor–Couette flow, there is a series of bifurcations from well-ordered to chaotic and ultimately to turbulent flow in Rayleigh–Benard convection. The nature of the transitions depends not only on the Rayleigh number  $Ra$  but also on a second dimensionless parameter called the *Prandtl number*  $Pr = \nu/\kappa$  which denotes the relative magnitudes of the viscous and thermal diffusivities and is a property of the fluid used in the experiments. Large Prandtl numbers are achieved in fluids with large viscosities; even water at room temperature has a Prandtl number of around 6. Pressurized gasses have Prandtl numbers of around 1. Small Prandtl numbers are achieved in flows of liquid metals (e. g., mercury or sodium) that are very good thermal conductors and flows in gasses which have a very small kinematic viscosity.

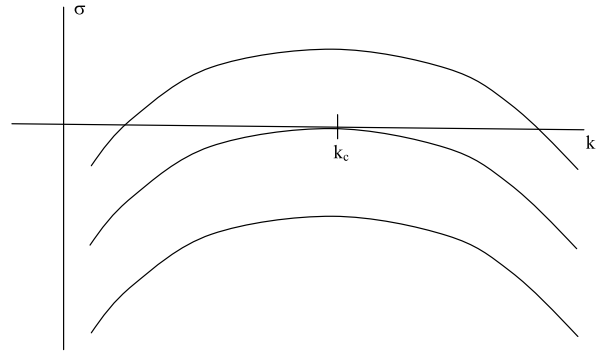
For fluids with small  $Pr$ , the first instability from steady (time-independent) convection rolls is an oscillatory instability, similar in many respects to the wavy vortex state for Taylor–Couette flow. The convection rolls form a snake-like instability along their axes, and this undulation in the rolls propagates in a direction parallel to the axes of the

convection rolls. A cross-section of the flow in this state would reveal a pattern of counter-rotating vortices that oscillates periodically in the lateral direction.

There are also several other instabilities that appear in Rayleigh–Bénard convection including: cross-roll instabilities in which a second pattern of convection rolls develops perpendicular to the first; a skew-varicose instability in which the width of convection rolls is modulated; a zig-zag instability which is exactly what its name implies, one in which the convection rolls form a zig-zag pattern; an instability referred to as the “Eckhaus instability” where pairs of convection rolls disappear, allowing other rolls present to grow, increasing the wavelength of the pattern; and spiral convection patterns. Defects can also appear in convection patterns where either two or more convection rolls meet at a point. The motion of a defect parallel to convection rolls can result in the addition or removal of a pair of convection rolls in the flow.

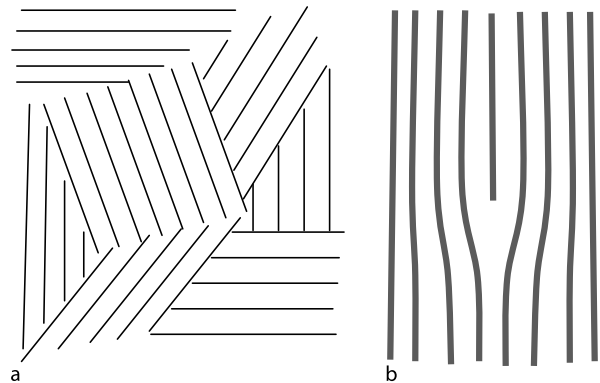
One way of understanding this behavior is to consider the stability of convection rolls near onset as a function of their wavelength. A common way of analyzing stability is via a “linear stability” argument. One starts with a motionless state, and then assumes a perturbation of a periodic pattern of convection rolls on this motionless state. A growth/decay term is associated with this perturbation, which might be written in the form  $A_0 e^{\sigma t} f(k)$  where  $f(k)$  is the perturbation with wavenumber  $k$ , and  $\sigma$  is the growth exponent. The growth exponent  $\sigma$  is typically dependent on the value of  $k$  (Fig. 5). If  $\sigma < 0$  for all values of  $k$ , then all perturbations decay and the motionless state is stable and no convection rolls will develop; this is the case for  $Ra < Ra_c$ . For  $Ra = Ra_c$ ,  $\sigma = 0$  for one particular critical wavenumber  $k_c$ ; for  $Ra$  just slightly above  $Ra_c$ , perturbations with wavenumber  $k_c$  should be expected to grow and, for an infinite system (i. e., no boundaries), a pattern of convection rolls will appear with precisely that wavenumber. For  $Ra > Ra_c$ , there is a range of wavenumbers for which  $\sigma > 0$ .

The discussion above is idealistic; in realistic systems, boundary conditions have a significant effect on convection patterns. If the fluid layer is confined to a small-to-moderate rectangular box, then the rolls line up parallel to the shortest side of the box, as in Fig. 4. If the width of the box is an integer multiple of the critical wavelength  $2\pi/k_c$ , then linear stability correctly predicts convection with this wavelength for  $Ra$  just barely above  $Ra_c$ . If the fluid layer is orders of magnitude wider than the critical wavelength, multiple zones of parallel convection rolls form at random orientation (Fig. 6a); the zones are separated by *grain boundaries*. Over time, some zones grow while others shrink. Ultimately, one zone may end up



**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 5**

Example of a plot of growth exponent  $\sigma$  as a function of wavenumber  $k$ , typical of results of linear stability analysis. The *bottom curve* corresponds to a Reynolds or Rayleigh number below the critical value for instability;  $\sigma$  is negative for all wavenumbers, so all disturbances decay. The *middle curve* corresponds to the onset of instability. For  $Ra$  or  $Re$  just above this value,  $\sigma$  is positive for wavenumbers very close to the critical value  $k_c$ . For larger  $Re$  or  $Ra$  (*top curve*), disturbances over a range of wavenumbers are unstable



**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 6**

**a** Sketch of zones of parallel convection rolls, separated by grain boundaries. The *lines* represent boundaries between adjacent rolls, as viewed from above (similar to dashed and dotted lines in Fig. 4). **b** Sketch of a defect in a convection pattern

dominating, with a single pattern of parallel convection rolls covering the entire system.

If contained in a circular container, a pattern of parallel convection rolls cannot be achieved [9], because convection rolls tend to line up perpendicular to the side walls. A theory by Swift and Hohenberg [10] defines a free-energy functional – an effective energy that depends on the functional form of the convection pattern; the pattern evolves to minimize this free energy. The free energy is

increased by convection rolls that are parallel to the side walls, by bends in convection rolls, and by defects and grain boundaries in the patterns. Near onset (i. e., for  $Ra$  just slightly over  $Ra_c$ ), a complicated convection pattern forms to minimize this free energy functional (Fig. 7a). The patterns that form have local wavenumbers that can not all be at the critical value  $k_c$ ; consequently, the pattern may continually evolve in time, in contradiction to the typical case in which convection with  $Ra$  just above  $Ra_c$  is time-independent.

For large enough  $Ra$ , convective flows become turbulent. Two types of turbulent flows are often discussed. At lower (but still large)  $Ra$ , complicated convection patterns form with numerous defects. The defects move around significantly in the pattern, resulting in continuous variation in the pattern. This regime is referred to as *weak turbulence* or *defect-mediated turbulence*. For very high  $Ra$ , the convection pattern breaks up completely and a regime of *strong turbulence* is found. A side view of convection in the strongly turbulent regime (Fig. 7b) shows that the flow is divided into three regions. Most of the temperature gradient is concentrated into thin boundary layers at the top and bottom of the fluid, while the fluid is well-mixed at roughly the average temperature in the large central region of the flow. The boundary layers are not quiescent; rather, they grow in thickness as heat flows into (for the lower boundary layer) and out of (for the upper boundary layer) them, eventually becoming unstable and forming plumes that erupt, carrying blobs of warm or cold fluid into the center region. The eruption of these plumes both acts to drive the turbulent flow and also transport heat from the bottom to the top of the system.

### Bénard–Marangoni Convection

If a fluid layer with a free surface is heated from below, thermal convection can be driven by variations in surface tension at the surface. This form of convection is referred to as Bénard–Marangoni convection and is usually characterized by a pattern of hexagons. Since the fluid is heated from below, it is unstably stratified. However, the main force driving the flow isn't buoyancy; rather, the flow is driven by gradients in the surface tension at the free surface. An instability occurs in which small variations in the temperature arise on the surface. The surface tension depends on the temperature; consequently, temperature gradients result in surface tension gradients that drive flows along the surface. Regions in which the fluid converges on the surface well downward, and upwelling of fluid occurs below regions in which the fluid diverges on the surface.

### Transport and Mixing

The manner in which an impurity is mixed within a fluid is an issue with significant practical importance for a wide range of scientific and engineering applications. Mixing of a *passive* impurity (one that follows fluid elements in the flow and whose distribution does not affect the flow itself) is governed by the advection-diffusion equation:

$$\frac{\partial c}{\partial t} = -\vec{u} \cdot \vec{\nabla} c + D \nabla^2 c,$$

where  $c$  is a scalar concentration field,  $\vec{u}$  is the velocity field, and  $D$  is the molecular diffusion coefficient. This equation can be written in non-dimensional form by scaling the velocity field by a characteristic flow velocity  $U$ , distance by a characteristic length  $L$ , and time by the advection timescale  $L/U$ :

$$\frac{\partial c}{\partial t'} = -\vec{u}' \cdot \vec{\nabla}' c + \frac{1}{Pe} \nabla'^2 c,$$

where the Peclet number  $Pe = UL/D$  characterizes the relative strength of advective to diffusive mixing in the flow. In the limit of no flow ( $Pe \rightarrow 0$ ), mixing is entirely via molecular diffusion, caused by Brownian motion of individual tracer particles in the impurity. Molecular diffusion is a slow mixing process; the variance  $\sigma^2(t)$  of a distribution grows linearly with time:  $\sigma^2(t) = 2Dt$ .

The presence of a fluid flow dramatically changes the mixing. An impurity is advected along with fluid elements in the flow, but can diffuse away from these deterministic fluid trajectories. Advection of an individual tracer is determined from equations of motion derived from the velocity field  $\vec{u} = (u_x, u_y, u_z)$ :

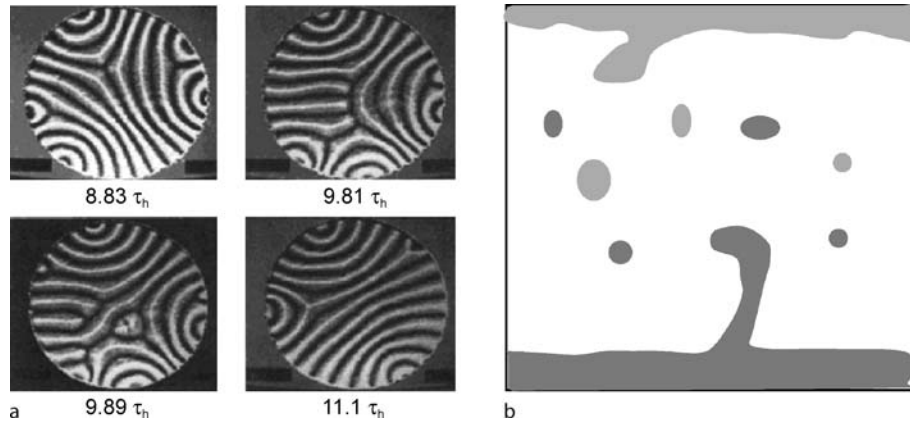
$$dx/dt = u_x$$

$$dy/dt = u_y$$

$$dz/dt = u_z.$$

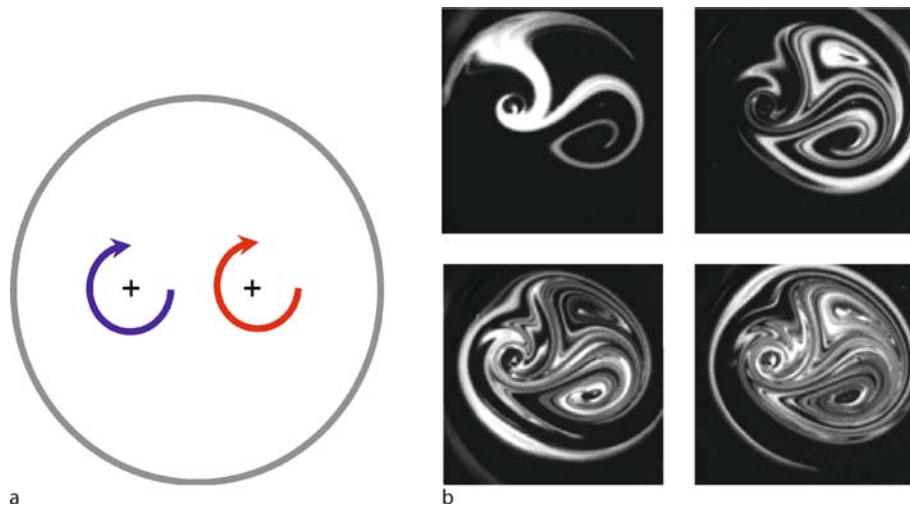
Trajectories are determined by integrating these equations of motion. If the flow is two-dimensional and time-independent ( $\vec{u} = \vec{u}(x, y)$  and  $u_z = 0$ ), then tracer trajectories are *integrable* (well-ordered). For two-dimensional, time-periodic or three-dimensional, time-independent flows, however, the equations of motion in general are not integrable, and trajectories are often chaotic, with nearby tracers separating roughly exponentially in time (sensitive dependence on initial conditions).

The most well-known flow exhibiting *chaotic advection* is the blinking vortex flow proposed by Hassan Aref [11]. The flow is shown in Fig. 8a; the fluid alternates periodically between circling around the left vortex



Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 7

**a** Images of Rayleigh–Bénard convection in a circular container near onset, viewed from above (from [9]). **b** Sketch of strongly-turbulent convection from side, showing the formation of thermal boundary layers at the bottom and top that erupt, sending plumes of hot and cold fluid into the middle region



Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 8

**a** Sketch of blinking vortex flow. Fluid alternates between circling around left and right vortex centers. **b** Mixing of dye in blinking vortex flow. Images are separated in time by one period of oscillation each (from [12])

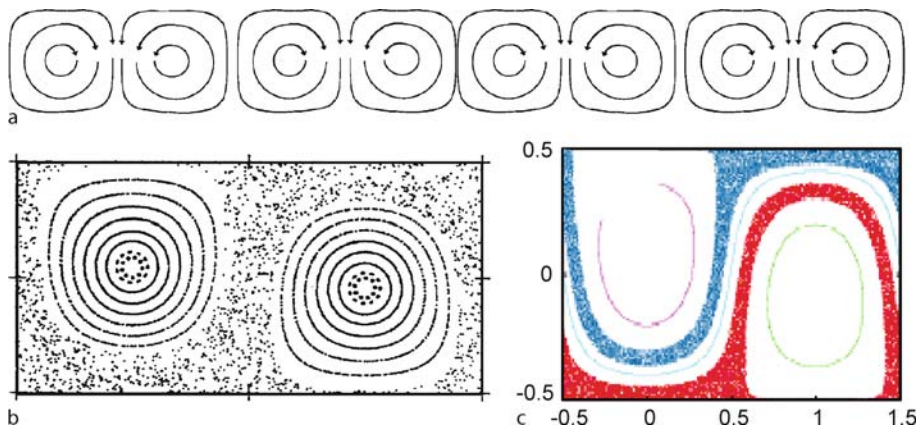
and around the right vortex. Advection of a blob of a passive dye is shown in Fig. 8b [12]. Mixing is associated with a repeated process of stretching and folding of tendrils of the impurity. Stretching and folding behavior such as this is typical of chaotic mixing.

Transport on scales larger than typical length scales of the flow can be illustrated with the alternating vortex chain [13], shown in Fig. 9. If the (two-dimensional) flow is time-independent, tracers follow ordered, closed trajectories within individual vortices. The presence of molecular diffusion allows for tracers to diffuse from one vortex to the next. The combination of advection around and diffusion between the vortices results in enhanced trans-

port whose variance grows linearly in time:  $\sigma^2(t) = 2D^*t$ , where  $D^*$  is the *enhanced diffusion coefficient*.

The simplest time dependence for the alternating vortex chain is a simple lateral oscillation of the entire chain, similar to time dependence found in Taylor–Couette and Rayleigh–Benard flows for intermediate values of  $Re$  and  $Ra$ . A passive tracer moving in this flow circles around an individual vortex, but if it is near a separatrix (boundary) between vortices, it can cross between vortices. Over time, the tracer alternates between circling within and crossing between vortices in an erratic manner. The trajectories are rigorously chaotic in the sense that nearby trajectories separate roughly exponential as a function of time.





**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 9**

**a** Sketch of alternating vortex chain. **b** Poincaré section for oscillating vortex chain, showing ordered regions of transport in vortex cores and chaotic region around and between vortices. **c** Poincaré section for oscillating and drifting vortex chain (from [14])

Not all tracers in the flow follow chaotic trajectories. Tracers near the center of a vortex remain trapped within that vortex, following a trajectory that is well-ordered and *not* sensitive to initial conditions. The separation of the system into both ordered and chaotic regions of transport can be seen easily by plotting a *Poincaré section* (Fig. 9b), a plot that shows the location of one tracer (or more) once every period of the flow. A tracer that starts in the chaotic region will eventually explore the entire chaotic region, which shows up on the Poincaré section as a pattern of dots. A chaotic trajectory can never cross into a region with ordered trajectories; consequently, the ordered regions show up as white “islands” on a Poincaré section unless tracers start within the ordered regions, revealing a pattern of closed curves, as seen in Fig. 9a.

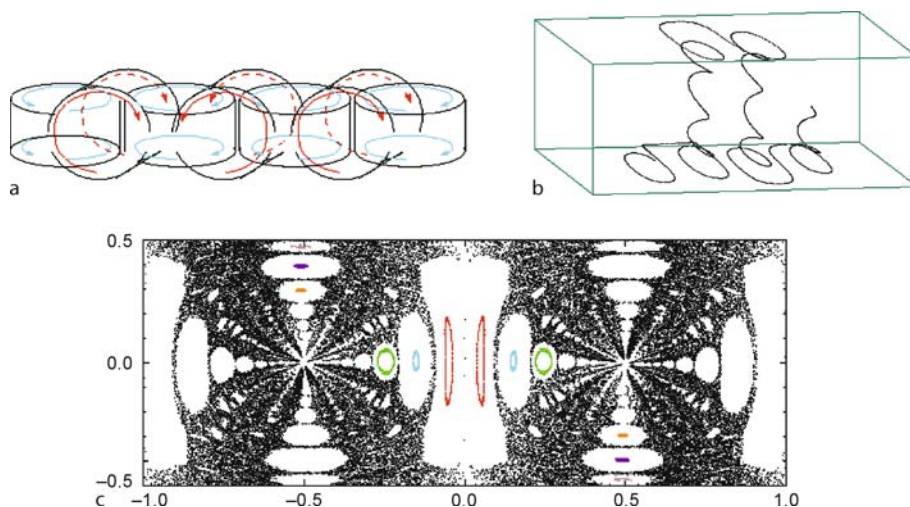
The boundaries between ordered and chaotic regions – referred to as *Kolmogorov–Arnold–Moser (KAM) invariant surfaces* – act as barriers against transport and mixing in the flow. KAM barriers are very common in laminar flows and also show up in many flows found in nature, even turbulent ones. Examples of transport barriers include the ozone hole over Antarctica, Gulf Stream rings in the Atlantic Ocean, and coherent vortices (called *meddies*) originating in the Mediterranean Sea that cross the Atlantic and maintain the salinity and biological constituency found in the Mediterranean for months and even years. Another well-known example of a transport barrier is the Great Red Spot of Jupiter, a large coherent patch of vorticity in a very turbulent atmosphere that clearly acts as a transport barrier, evidenced by the fact that it has maintained its distinct color for centuries.

Chaotic advection results in significant enhancements in long-range transport. In many cases, enhanced transport is diffusive with a variance that grows linearly in time.

However, *anomalous diffusion* is also possible where the variance grows as a power law in time with a growth exponent different from 1:  $\sigma^2(t) \sim t^\gamma$  with  $\gamma \neq 1$ . Transport with  $\gamma < 1$  is called *subdiffusion* and transport with  $\gamma > 1$  is called *superdiffusion*. The presence of islands of ordered trajectories is often associated with subdiffusive transport; tracers in the chaotic region stick to the outside of these islands and possibly diffuse inward due to Brownian motion. Superdiffusion is associated with *Lévy flight* trajectories where tracers jump large distances between regions in the flow. An example of a flow with superdiffusion is the alternating vortex chain with both oscillatory and drifting time dependence [14] (Fig. 9a). A Poincaré section for transport in this flow is shown in Fig. 9c. Two distinct chaotic regions are apparent, and in addition to ordered islands in the vortices, there is also a snake-like region of ordered trajectories that wind around the vortices. A tracer in the chaotic region in this flow alternately sticks to islands (remaining temporarily confined to a vortex) and to the snake region (traveling rapidly along the vortex chain).

Chaotic advection is also possible in three-dimensional flows, even those that are time-independent. An example of a such a flow is shown in Fig. 10 – a chain of alternating vortices superposed with a second chain which is shifted by half of a vortex width and is oriented with its vortex axes perpendicular to those of the first chain [15]. Trajectories of tracers in this flow are shown in Fig. 10b. A Poincaré section (Fig. 10c) again shows coexistence of both ordered and chaotic regions of transport in the flow.

There are numerous applications of chaotic mixing. Chemical engineers have already begun to develop and market devices that take advantage of chaos to enhance mixing in low-Reynolds number flows. Chaotic mixing



**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 10**

**a** Sketch of time-independent, nested vortex chain flow. **b** Example of a chaotic trajectory for the nested vortex flow. **c** Poincaré section at the mid-height of the flow, showing both ordered and chaotic regions (from [15])

is particularly useful for mixing of high-viscosity fluids for which turbulent mixing would be difficult and/or tremendously expensive energetically. Chaotic mixing is also finding significant applications in the expanding field of microfluidic devices, so-called “factories-on-a-chip,” again which involve very small Reynolds numbers due to the small length-scales of these systems.

### Pattern Formation in Reaction-Diffusion and Advection-Reaction-Diffusion Systems

Another class of nonlinear, pattern-forming systems is the well-known reaction-diffusion (RD) problem. The reaction is a process in which one or more species somehow changes into something different. Examples include chemical reactions, biological interaction of some sort (such as the interaction of a disease with a population of people or animals or a predator-prey system), and a wide range of phase transitions. In an extended, non-flowing system, the reaction typically occurs at different times in different locations in the system, resulting in the formation of patterns. The key to the formation of these patterns is the interaction between the reaction at local regions in the system and molecular diffusion which (for a non-flowing system) is the only means of communication between different regions.

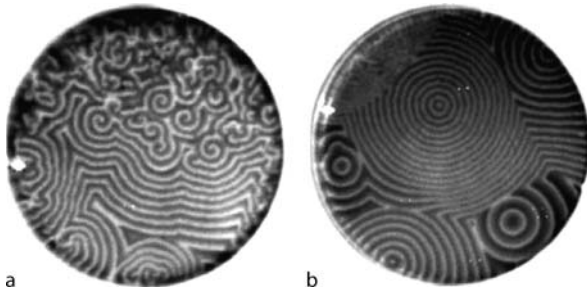
A general form of the RD equation can be written as follows:

$$\frac{\partial c}{\partial t} = f(c) + D\nabla^2 c,$$

where  $c$  is a concentration field of some chemical, biological or physical species, and the first term on the right denotes some sort of reaction. The details of the reaction term itself vary from system to system; however, some common behavior is found in a wide range of RD systems, independent of the details of this reaction term.

The field of reaction-diffusion studies was enhanced significantly by the discovery of the oscillatory *Belousov-Zhabotinsky* (BZ) chemical reaction. Discovered in the 1950s, this reaction is well-known to have a pH that oscillates for hours when well-mixed in a beaker or a flask. When catalyzed with a pH-indicator such as ferroin, the chemicals can be observed to alternate back and forth between a red and blue color during this period before the reaction finally runs down. The oscillation is typically nearly-periodic, but conditions have been found in which chaotic oscillations can be found in the BZ reaction.

If the chemicals for the BZ reaction are in a shallow petrie dish with no flow (or in a gel or fritted disk that inhibits a flow), spiral and/or target patterns form, as seen in Fig. 11. Patterns similar to these have been observed in a wide range of physical, chemical and biological systems. Examples include (a) spiral waves of electrical activity in the heart which act as pacemakers; breakdown of these spiral waves are associated with cardiac fibrillation; (b) spiral waves of “spreading depression” in the brain that are associated with migraine headaches; (c) spiral patterns found in developing embryos which may be partially responsible for morphogenesis in which different cells in the embryo develop different roles in the growing organism; and (d) spiral and target patterns in populations of slime molds



**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 11**

Examples of reaction-diffusion patterns formed by the Belousov–Zhabotinsky chemical reaction in a thin layer in a petrie dish

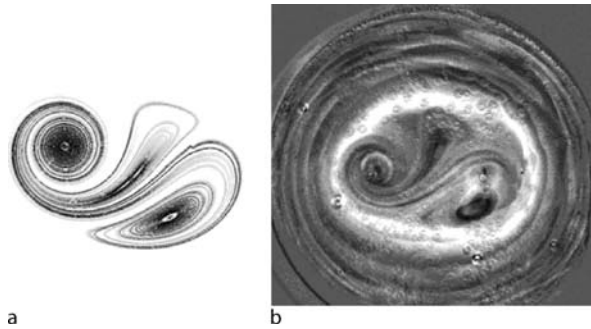
in a stagnant system. Because of the similarity of these patterns to those found in the BZ system, there has been a significant amount of research into the properties of the BZ system, which is considered to be a paradigm of RD systems. Reaction-diffusion dynamics also explain the propagation of fronts in stagnant systems, e. g., the spreading of a fire in the absence of any winds or the growth of a solid in the absence of fluid flows.

In both natural and industrial systems, reacting systems are rarely stagnant; flows dramatically affect the mixing and therefore the communication between different parts of the system. The more general *advection-reaction-diffusion* (ARD) problem is represented by the following equation:

$$\frac{\partial c}{\partial t} = -\vec{u} \cdot \vec{\nabla} c + f(c) + D \nabla^2 c$$

where the additional term on the right denotes advection of the relevant species by the fluid flow. Advection-reaction-diffusion dynamics are relevant for pattern formation and front propagation in a wide range of systems, including marine ecological systems (e. g., algae and phytoplankton blooms), microfluidic chemical and biological processing and diagnostics, forest fires in the presence of winds, and the propagation of disease in a moving population.

If the mixing is chaotic, it has been shown [16] that ARD patterns reflect the structures that characterize chaotic mixing in the flow. An example [12] is the BZ reaction in a blinking vortex flow (Fig. 8) in which fluid circulates alternately (and periodically) between two separate vortex centers. As seen in Fig. 12, except for very weak mixing, the patterns associated with chaotic mixing end up dominating the pattern-formation process for the ARD system as well [12]. Ideas such as this have been used to explain patterns of populations of marine organisms in both the Gulf of Mexico and in the northern Atlantic Ocean.



**Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 12**

**a** Simulation of mixing field for blinking vortex flow. **b** Advection-reaction-diffusion pattern for Belousov–Zhabotinsky chemical reaction in the same flow (from [12])

### Other Examples of Pattern Forming Systems

Patterns similar to those discussed above are found in a wide range of systems spanning all fields of science and engineering. Some examples are as follows:

- Electrohydrodynamic convection in flows of nematic liquid crystals.
- Wave patterns on the surface of a fluid oscillated periodically in the vertical direction.
- Vibrating granular systems.
- Bubble froths.
- Ferrofluids – labyrinth patterns.

This is just a small sample of the many nonlinear, pattern-forming systems that have been studied during the past few decades.

### Future Directions

The study of nonlinear systems is an on-going and continually evolving field. Nonlinear dynamics span a wide range of fields of study, including all of the fields of sciences and engineering, as well as mathematics, medicine, economics and even some fields of social sciences. Not only are there numerous on-going studies of the basic mathematical and scientific behavior of nonlinear systems, but there are also many applications that are being developed, based on the principles of nonlinear systems.

### Bibliography

#### Primary Literature

1. Landau L (1944) CR (Dokl) Acad Sci URSS 44:311
2. Gollub JP, Swinney HL (1975) Phys Rev Lett 35:927
3. Ruelle D, Takens F (1971) Commun Math Phys 20:167

4. Marcus PS (1988) *Nature* 331:693; Sommeria J, Meyers SD, Swinney HL (1988) *Nature* 331:689
5. Kalliroscope can be obtained from Kalliroscope Corporation. Groton, 978-448-6302. [www.kalliroscope.com](http://www.kalliroscope.com)
6. Fenstermacher PR, Swinney HL, Gollub JP (1979) *J Fluid Mech* 94:103
7. Andereck CD, Liu SS, Swinney HL (1986) *J Fluid Mech* 164:155
8. Andereck CD, Dickman R, Swinney HL (1983) *Phys Fluids* 26:1395
9. Heutmaker MS, Fraenkel PN, Gollub JP (1985) *Phys Rev Lett* 54:1369
10. Swift J, Hohenberg PC (1977) *Phys Rev A* 15:319
11. Aref H (1984) *J Fluid Mech* 143:1
12. Nugent CR, Quarles WM, Solomon TH (2004) *Phys Rev Lett* 93:218301
13. Solomon TH, Gollub JP (1988) *Phys Fluids* 31:1372; *Phys Rev A* 38:6280; Solomon TH, Tomas S, Warner JL (1996) *Phys Rev Lett* 77:2682
14. Paoletti MS, Nugent CR, Solomon TH (2006) *Phys Rev Lett* 96:124101
15. Fogleman MA, Fawcett MJ, Solomon TH (2001) *Phys Rev E* 63:020101(R)
16. Tel T, de Moura A, Grebogi C, Karolyi G (2005) Chemical and biological activity in open flows: A dynamical system approach. *Phys Rep* 413:91

### Books and Reviews

- Baker GL, Gollub JP (1990) *Chaotic dynamics: An introduction*. Cambridge University Press, Cambridge
- Ben-Avraham D, Havlin S (2000) *Diffusion and reactions in fractals and disordered systems*. Cambridge University Press, Cambridge
- Cross MC, Hohenberg PC (1993) Pattern-Formation outside of equilibrium. *Rev Mod Phys* 65:851
- Ott E (2002) *Chaos in dynamical systems*, 2nd edn. Cambridge University Press, Cambridge
- Tritton DJ (1988) *Physical fluid dynamics*, 2nd edn. Clarendon Press, Oxford
- Winfree AT (1980) *The geometry of biological time*. Springer, New York
- Grindrod P (1996) *The theory and applications of reaction-diffusion equations: patterns and waves*. Clarendon Press, Oxford

## Non-linear Internal Waves

MOUSTAFA S. ABOU-DINA, MOHAMED A. HELAL  
Department of Mathematics, Faculty of Science,  
Cairo University, Giza, Egypt

### Article Outline

[Glossary](#)  
[Definition of the Subject](#)  
[Introduction](#)  
[Problem and Frame of Reference](#)  
[Notation](#)  
[Equations of Motion](#)

### The Shallow Water Theory

[Free Surface and Interface Elevations of Different Modes](#)  
[Secular Term](#)

[Multiple Scale Transformation of Variables](#)

[Derivation of the KdV Equation](#)

[Conclusions](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Nonlinear waves** Nonlinear waves are such waves which arise as solutions of the nonlinear mathematical models simulating physical phenomena in fluids.

**Shallow water** Shallow water means water waves for which the ratio between the amplitude and the wave length is relatively small. The linear theory of motion is inadequate for the description of shallow water waves.

**Internal waves** Internal waves are gravity waves that oscillate within a fluid medium. They arise from perturbations to hydrostatic equilibrium, where balance is maintained between the force of gravity and the buoyant restoring force. A simple example is a wave propagating on the interface between two fluids of different densities, such as oil and water. Internal waves typically have much lower frequencies and higher amplitudes than surface gravity waves because the density differences (and therefore the restoring forces) within a fluid are usually much smaller than the density of the fluid itself.

**Pycnocline** Pycnocline is a rapid change in water density with depth. In freshwater environments, such as lakes, this density change is primarily caused by water temperature, while in seawater environments such as oceans the density change may be caused by changes in water temperature and/or salinity.

**Solitary waves** Solitary waves are localized traveling waves, which asymptotically tend to zero at large distances.

**Solitons** Solitons are waves which appear as a result of a balance between a weakly nonlinear convection and a linear dispersion. The solitons are localized highly stable waves that retain their identity (shape and speed) upon interaction, and resemble particle like behavior. In the case of a collision, solitons undergo a phase shift.

**Baroclinic fluid** Baroclinic fluid is such a fluid for which the density depends on both the temperature and the pressure.

In atmospheric terms, the baroclinic areas are generally found in the mid-latitude/polar regions.



**Barotropic fluid** Barotropic fluid is such a fluid for which the density depends only on the pressure.

In atmospheric terms, the barotropic zones of the Earth are generally found in the central latitudes, or tropics.

## Definition of the Subject

The objective of the present work is to study the generation and propagation of nonlinear internal waves in the frame of the shallow water theory. These waves are generated inside a stratified fluid occupying a semi infinite channel of finite and constant depth by a wave maker situated in motion at the finite extremity of the channel. A distortion process is carried out to the variables and the nonlinear equations of the problem using a certain small parameter characterizing the motion of the wave maker and double series representations for the unknown functions is introduced. This procedure leads to a solution of the problem including a secular term, vanishing at the position of the wave maker. This inconvenient result is remedied using a multiple scale transformation of variables and it is shown that the free surface and the interface elevations satisfy the well known KdV equation. The initial conditions necessary for the solution of the KdV equations are obtained from the results of the first procedure.

## Introduction

The work on internal gravity waves has been started in the middle of the twentieth century by Keulegan [21] followed by Long [25], where they applied Boussinesq's and Rayleigh's techniques, respectively, on free surface waves to the problem of internal waves.

In physical oceanography, fluid density varies with the depth due to the salinity and the temperature of the fluid [23]. Several works, dealing with variable density, have been worked out (e. g. [2,4,15,19,20,31,32], ...).

The geophysical problem of the propagation waves in stratified fluid is very important in physical oceanography. The physical problem is simulated by a suitable model of a free-surface and interface fluid flow over a horizontal bottom or over a certain topography.

The fluid in the ocean may be considered as a stratified one according to the variation of the salinity and temperature, due to the weather, with respect to the vertical coordinate normal to the surface. The mathematical model of a two-fluid system (stratified fluid) is a good description to simulate wave motions in physical applications.

Based on results of many field observations, it is found that the fluid's density changes rapidly within the regions of pycnocline.

Nonlinear theoretical models for waves in a two-fluid system have been established with various restrictions on length scales. Among these, Benjamin [6] derived the KdV equation for thin layers in comparison with the wave length. Later, Benjamin [7], Davis and Acrivos [14] and Ono [30] constructed the BO equation under the assumptions of a thin upper layer and an infinitely deep lower layer. Kubota et al. [22] and Choi and Camassa [11,12] carried out a series of investigations to derive model equations for weakly and strongly nonlinear wave propagation in stratified fluids. The upper boundary is allowed to be either free or rigid. The assumptions they stated are as follows:

- 1) the wavelength of the interfacial wave is long.
- 2) the upper layer is thin.
- 3) no depth restriction is made on the lower layer.

The effect of a submerged obstacle on an incident wave inside an ideal two-layer stratified shallow water has been studied in two dimensions by Abou-Dina and Helal [3] by applying the shallow water approximation theory.

In 1995, Pinettes, Renouard and Germain presented a detailed analytical and experimental study of the oblique passing of a solitary wave over a shelf in a two-layer fluid.

Later, Barthelemy, Kabbaj and Germain [5] applied the WKB technique to investigate theoretically the scattering of a surface long wave by a step bottom in a two-layer fluid. This method enabled them to overcome the main inconsistency of the shallow-water theory in the presence of obstacles.

Matsuno [28] attempted to unify the KdV, BO and ILW models using the assumption of small wave steepness. Lynett and Liu [26] assumed a small density difference and derived a set of model equations.

Recently, Craig et al. [13] presented a Hamiltonian perturbation theory for the long-wave limit, and provided a uniform treatment of various long-wave models (KdV, BO and ILW models). Their formulation is shown to be very effective for perturbation calculations and represents a basis for numerical simulations. For a single-fluid system, in the last two decades, the traditional Boussinesq equations have been improved to extend their applicability from shallow water to deep water (see for example: Nwogu [29], Chen and Liu [10], Wei et al. [36], Madsen and Schaffer [27], Gobbi et al. [18]).

Most recently, Liu et al. [24] published an excellent study on the essential properties of Boussinesq equations for internal and surface waves in a two-fluid system.

Most studies applied the technique of the Padé approximate to allow a much higher-order accuracy without increasing the order of the derivatives.



Although the theoretical work for surface waves are excessive, studies of internal waves based on Boussinesq equations are still lack.

In the present work, and for a better description of the physical problem, we shall investigate, in the frame of the shallow water theory, the effect of the bounded motion of a vertical plane wave maker on the generation and propagation of waves inside a stratified fluid. The wave maker is situated at the finite end of a semi-infinite channel of constant depth occupied by two non immiscible fluid layers of different constant densities. The problem considered here, is supposed to be a two dimensional one with an irrotational motion, the fluid layers are taken ideal and the surface tension is neglected.

Double series representation for the unknown functions, in terms of a certain small parameter characterizing the motion of the wave maker is used. The first few orders of approximations obtained following this technique indicates the presence of a secular term increasing indefinitely in going away from the wave maker and vanishing at the position of this later. To overcome this physically unacceptable result, a multiple scale transformation of variables is carried out and a single series representation, for the unknown functions, is proposed. This technique leads to solutions up to the fifth order of approximations free from secular terms and shows that the generation and propagation processes of internal waves are governed by the well known KdV equation. The initial conditions needed for the solution of this nonlinear partial differential equation are obtained from the results of the preceding method at the position of the wave maker where the solution is free from secular terms. On the light of this result it is expected to obtain an internal solitonic waves.

Problem and Frame of Reference

A vertical plane wave-maker is situated at the finite extremity of a semi-infinite channel with a constant depth. The channel is occupied by two layers of immiscible and inviscid liquids of constant densities. The wave-maker is set in motion generating waves which propagate, inside both fluid layers, towards the down-stream extremity of the channel. The required is to calculate the produced waves and also to determine the motion that should be assigned to the wave maker in order to generate internal waves only in the channel.

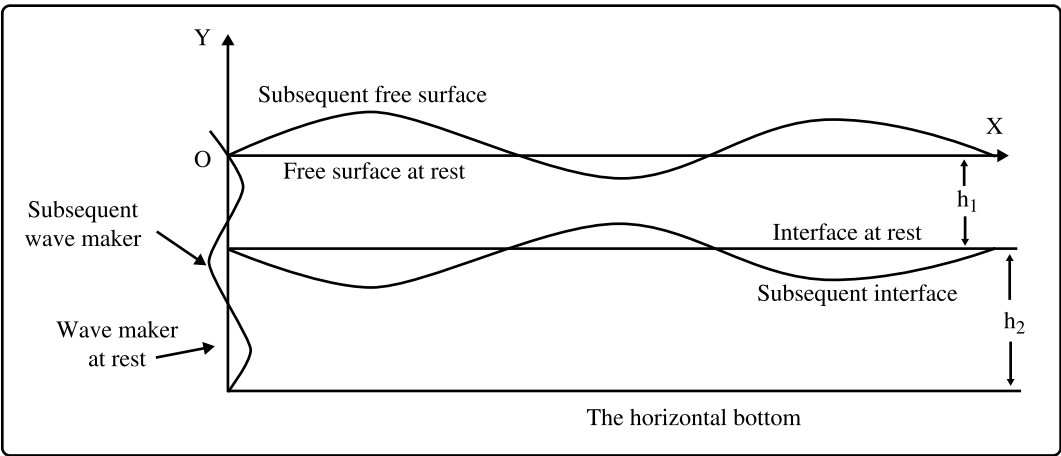
In the mathematical model, the problem is assumed to be a two dimensional one, the flow is considered irrotational and the free surface and the interface are assumed to remain always near their positions at rest. Furthermore, the motion of the wave maker is considered slow and bounded and the problem is studied in the frame of the non-linear shallow water theory of motion.

A fixed rectangular system of reference is used for the description of the motion of the fluid. The origin  $O$  is taken in the free surface at rest with horizontal  $x$ -axis pointing along the direction of propagation of waves. The  $y$ -axis is directed vertically upwards (Fig. 1).

Notation

The following notation is used throughout this paper:

$g$	the acceleration of gravity
$h_1, h_2 (h = h_1 + h_2)$	the constant depths of the upper and lower layers of the fluid, respectively
$P(x, y)$	the pressure
$S$	the stocks parameter



Non-linear Internal Waves, Figure 1  
Problem and frame of reference

$t$	the time
$D_x = \varepsilon f(y, \varepsilon t)$	the horizontal displacement of the wave maker ( $= \varepsilon f^{(1)}(y, \varepsilon t)$ in the upper layer and $\varepsilon f^{(2)}(y, \varepsilon t)$ in the lower layer)
$\Phi^{(j)}(x, y, t)$	the velocity potentials: $j = 1, 2$ for the upper and lower layers, respectively
$y = \eta^{(1)}(x, t)$	the equation of the free surface
$y = -h_1 + \eta^{(2)}(x, t)$	the equation of the interface
$(x, y)$	Cartesian coordinates of a point
$\delta_{n,m}$	the Kronecker delta = 1 if $n = m$ ; 0 if $n \neq m$
$\varepsilon$	a small parameter
$\rho_1, \rho_2$	the constant densities of the upper and lower layers, respectively ( $\rho_1 < \rho_2$ )
$D_z$	an operator of total differentiation w.r.t. $z$
$\partial_{x,y,\dots,z}$	an operator of partial differentiation w.r.t. $x, y, \dots$ , and $z$

### Superscripts

1	upper layer
2	lower layer
', ''	first and second derivatives w.r.t. the argument of the superscripted function

### Subscripts

e	the barotropic (external) mode
i	the baroclinic (internal) mode

### Equations of Motion

The general system of equations and simplifying conditions describing water wave motion are expressed in terms of the velocity potential in the case of irrotational motion for both homogeneous and stratified fluids (Boussinesq [9], Stoker [33], Wehausen and Laitone [35]).

As in Abou-Dina and Helal [2,3,4], we start with the set of distorted variables:  $\hat{x}, \hat{y}, \hat{t}$  defined in terms of  $x, y, t$  as:

$$\hat{x} = \varepsilon x, \quad \hat{y} = y, \quad \hat{t} = \varepsilon t, \quad (1)$$

where  $\varepsilon$  is a small parameter.

Let us denote by  $\hat{\Phi}^{(j)}, \hat{P}^{(j)}$  and  $\hat{\eta}^{(j)}$  the functions  $\Phi^{(j)}(\hat{x}/\varepsilon, \hat{y}, \hat{t}/\varepsilon)$ ,  $P^{(j)}(\hat{x}/\varepsilon, \hat{y}, \hat{t}/\varepsilon)$  and  $\eta^{(j)}(\hat{x}/\varepsilon, \hat{t}/\varepsilon)$  respectively.

According to the physical and simplifying conditions adopted here, the system of equations and conditions governing the problem is written in terms of the distorted variables  $\hat{x}, \hat{y}, \hat{t}$  as follows (Abou-Dina and Helal [2,3,4], Abou-Dina and Hassan [1]):

(I) In the fluid layers with constant densities:

$$\varepsilon^2 \partial_{\hat{x}\hat{x}} \hat{\Phi}^{(j)} + \partial_{\hat{y}\hat{y}} \hat{\Phi}^{(j)} = 0, \quad (2)$$

$$\hat{P}^{(j)}(\hat{x}, \hat{y}, \hat{t}) = \rho_j \left[ \varepsilon \partial_{\hat{t}} \hat{\Phi}^{(j)} + \frac{1}{2} \left( \varepsilon^2 \left( \partial_{\hat{x}} \hat{\Phi}^{(j)} \right)^2 + \left( \partial_{\hat{y}} \hat{\Phi}^{(j)} \right)^2 \right) + g\hat{y} \right], \quad (3)$$

where  $j$  hereafter stands for both 1 and 2.

(II) On the free surface which is impermeable and isobaric:

$$\partial_{\hat{y}} \hat{\Phi}^{(1)} = \varepsilon^2 \partial_{\hat{x}} \hat{\eta}^{(1)} \partial_{\hat{x}} \hat{\Phi}^{(1)} + \varepsilon \partial_{\hat{t}} \hat{\eta}^{(1)} \quad \text{at } \hat{y} = \hat{\eta}^{(1)}(\hat{x}, \hat{t}), \quad (4)$$

$$\varepsilon \partial_{\hat{t}} \hat{\Phi}^{(1)} + \frac{1}{2} \left( \varepsilon^2 \left( \partial_{\hat{x}} \hat{\Phi}^{(1)} \right)^2 + \left( \partial_{\hat{y}} \hat{\Phi}^{(1)} \right)^2 \right) + g\hat{y} = 0 \quad \text{at } \hat{y} = \hat{\eta}^{(1)}(\hat{x}, \hat{t}). \quad (5)$$

(III) At the interface, the impermeability gives:

$$\partial_{\hat{y}} \hat{\Phi}^{(j)} = \varepsilon^2 \partial_{\hat{x}} \hat{\eta}^{(2)} \partial_{\hat{x}} \hat{\Phi}^{(j)} + \varepsilon \partial_{\hat{t}} \hat{\eta}^{(2)} \quad \text{at } \hat{y} = -h_1 + \hat{\eta}^{(2)}(\hat{x}, \hat{t}), \quad (6)$$

Also the compatibility condition for the pressure across this boundary implies:

$$\begin{aligned} & \rho_1 \left[ g \left( h_1 + \hat{\eta}^{(2)} \right) + \varepsilon \partial_{\hat{t}} \hat{\Phi}^{(1)} + \frac{1}{2} \left( \varepsilon^2 \left( \partial_{\hat{x}} \hat{\Phi}^{(1)} \right)^2 + \left( \partial_{\hat{y}} \hat{\Phi}^{(1)} \right)^2 \right) \right] \\ &= \rho_2 \left[ g \left( h_1 + \hat{\eta}^{(2)} \right) + \varepsilon \partial_{\hat{t}} \hat{\Phi}^{(2)} + \frac{1}{2} \left( \varepsilon^2 \left( \partial_{\hat{x}} \hat{\Phi}^{(2)} \right)^2 + \left( \partial_{\hat{y}} \hat{\Phi}^{(2)} \right)^2 \right) \right] \\ & \text{at } \hat{y} = -h_1 + \hat{\eta}^{(2)}(\hat{x}, \hat{t}). \end{aligned} \quad (7)$$

(IV) On the horizontal bottom:

$$\partial_{\hat{y}} \hat{\Phi}^{(2)} = 0 \quad \text{at } \hat{y} = -h. \quad (8)$$

(V) The radiation condition implies that no wave comes from infinity.

- (VI) The initial conditions: At the initial instant of time ( $\hat{t} = 0$ ), the fluid is at rest with horizontal free surface at  $\hat{y} = 0$  and interface at  $\hat{y} = -h_1$ . This initial condition assumes that the functions  $f^{(j)}(\hat{y}, \hat{t})$  have initial vanishing values and implies at  $\hat{t} = 0$  that

- inside the fluid layers

$$\hat{\Phi}^{(j)}(\hat{x}, \hat{y}, 0) = 0, \quad (9a)$$

$$\partial_{\hat{x}} \hat{\Phi}^{(j)}(\hat{x}, \hat{y}, 0) = 0, \quad (9b)$$

$$\partial_{\hat{y}} \hat{\Phi}^{(j)}(\hat{x}, \hat{y}, 0) = 0, \quad (9c)$$

- at the free surface

$$\hat{\eta}^{(1)}(\hat{x}, 0) = 0 \quad (9d)$$

and

- at the interface

$$\hat{\eta}^{(2)}(\hat{x}, 0) = 0, \quad (9e)$$

provided that

$$f^{(j)}(\hat{y}, 0) = 0 \quad \text{for } j = 1, 2. \quad (9f)$$

- (VII) On the wave maker:

$$\partial_{\hat{x}} \hat{\Phi}^{(j)}(\hat{x}, \hat{y}, \hat{t}) = \varepsilon \frac{\partial}{\partial \hat{t}} f^{(j)}(\hat{y}, \hat{t})$$

$$\text{at } \hat{x} = 0, \quad -h + h_2 \delta_{j,1} \leq \hat{y} \leq h_1 \delta_{j,2}. \quad (10)$$

to the system of equations and conditions governing the problem in the non-distorted (physical) space, we have to replace  $\varepsilon$  by unity and omit the hats (^) over the symbols in Eqs. (2) to (10).

### The Shallow Water Theory

In the frame work of the shallow water theory, the non-linear system of Eqs. (2) to (10) will be solved using the double series representation:

$$\hat{\Phi}^{(j)}(\hat{x}, \hat{y}, \hat{t}) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \varepsilon^n \exp\left(-\frac{m\pi\hat{x}}{h_j\varepsilon}\right) \times \hat{\Phi}_{n,m}^{(j)}(\hat{x}, \hat{y}, \hat{t}), \quad (11a)$$

$$\hat{\eta}^{(j)}(\hat{x}, \hat{t}) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \varepsilon^n \exp\left(-\frac{m\pi\hat{x}}{h_j\varepsilon}\right) \hat{\eta}_{n,m}^{(j)}(\hat{x}, \hat{t}). \quad (11b)$$

The validity of using the above representation has been studied by Germain [16,17]. The small parameter  $\varepsilon$  represents the ratio of the water depth to the wave length. Hence,  $S \geq 1$  and we are dealing with the non-linear acoustic analogy.

According to the theory under consideration, at each order ( $n, m$ ), the above system of equations and conditions must be verified.

For simplification in the following, the hats (^) over the symbols will be omitted.

### Verification of the Homogeneous Equations

The above procedure, when applied to the homogeneous Eqs. (2)–(8) along with the radiation condition (V), can lead after some manipulations to the following expression for the total velocity potential up to the second order of the small parameter  $\varepsilon$ :

$$\begin{aligned} \Phi^{(j)}(x, y, t) = & \varepsilon \left[ A^{(j)}(x - C_1 t) + B^{(j)}(x - C_2 t) \right] \\ & + \varepsilon^2 \left[ R^{(j)}(x - C_1 t) + S^{(j)}(x - C_2 t) \right] \\ & + \varepsilon^2 \sum_{m=1}^{\infty} A_m^{(j)}(t) \exp\left(-\frac{m\pi x}{h_j \varepsilon}\right) \\ & \times \cos\left(\frac{m\pi}{h_j}(y + \delta_{j2}h)\right) + O(\varepsilon^3), \end{aligned} \quad (12)$$

where,

$$C_j^2 = \frac{g}{2} \left\{ h - (-1)^j \sqrt{(h_1 - h_2)^2 + 4(\rho_1/\rho_2) h_1 h_2} \right\}. \quad (13a)$$

The functions  $A^{(j)}$ ,  $B^{(j)}$  and  $A_m^{(j)}$  are arbitrary functions to be determined, with:

$$A^{(1)}(x - C_1 t) = \gamma_1 A^{(2)}(x - C_1 t) \quad (13b)$$

$$B^{(1)}(x - C_2 t) = \gamma_2 B^{(2)}(x - C_2 t) \quad (13c)$$

and,

$$\gamma_j = \frac{g h_2}{C_j^2 - g h_1}. \quad (13d)$$

The functions  $R^{(j)}$  and  $S^{(j)}$  are arbitrary functions combined by similar relations to those in Eq. (12) in terms of  $A^{(j)}$  and  $B^{(j)}$  respectively.

Equations (11)–(13) indicate the existence of two different wave speeds  $C_1$  and  $C_2$  ( $C_1 > C_2$ ). Accordingly, we obtain local oscillations along with progressive waves of two different modes:

- (i) External (or barotropic) modes corresponding to the wave speed  $C_1$ .
- (ii) Internal (or baroclinic) modes corresponding to the wave speed  $C_2$ .

The corresponding expressions of the free surface and the interface are given, to the third order, by:

$$\eta^{(1)}(x, t) = -(\varepsilon/g) \partial_t \Phi^{(1)}(x, y, t) \quad \text{at } y = 0, \quad (14a)$$

$$\begin{aligned} \eta^{(2)}(x, t) &= \left( \frac{\varepsilon}{g(\rho_2 - \rho_1)} \right) \partial_t \\ &\times \left\{ \rho_1 \Phi^{(1)}(x, y, t) - \rho_2 \Phi^{(2)}(x, y, t) \right\} \quad \text{at } y = -h_1. \end{aligned} \quad (14b)$$

### Complete Determination of the Solution

To complete the determination of the velocity potentials  $\Phi^{(j)}(x, y, t)$ , we verify the initial conditions (9) and the condition on the wave maker (10) which yield, together with expression (12) for the velocity potential functions, the following relations:

- In the upper layer:

$$\begin{aligned} A^{(1)'}(-C_1 t) + B^{(1)'}(-C_2 t) - \frac{\pi}{h_1} \sum_{m=1}^{\infty} m A_m^{(1)}(t) \\ \times \cos\left(\frac{m\pi}{h_1} y\right) = \frac{\partial}{\partial t} f^{(1)}(y, t), \end{aligned} \quad (15a)$$

$$-h_1 \leq y \leq 0.$$

- In the lower layer:

$$\begin{aligned} A^{(2)'}(-C_1 t) + B^{(2)'}(-C_2 t) - \frac{\pi}{h_2} \sum_{m=1}^{\infty} (-1)^m m A_m^{(2)}(t) \\ \times \cos\left(\frac{m\pi}{h_2}(y + h_1)\right) = \frac{\partial}{\partial t} f^{(2)}(y, t), \end{aligned} \quad (15b)$$

$$-h \leq y \leq -h_1.$$

Relations (15a) and (15b) together with the initial conditions (9) lead, after some manipulations, to the following expressions:

$$\begin{aligned} A^{(2)}(x - C_1 t) &= \frac{C_1}{\gamma_1 - \gamma_2} \\ &\times \left[ \frac{\gamma_2}{h_2} \int_{-h}^{-h_1} f^{(2)}\left(y, -\frac{x - C_1 t}{C_1}\right) dy \right. \\ &\left. - \frac{1}{h_1} \int_{-h_1}^0 f^{(1)}\left(y, -\frac{x - C_1 t}{C_1}\right) dy \right] \end{aligned} \quad (16a)$$

$$\begin{aligned} B^{(2)}(x - C_2 t) &= \frac{C_2}{\gamma_2 - \gamma_1} \\ &\times \left[ \frac{\gamma_1}{h_2} \int_{-h}^{-h_1} f^{(2)}\left(y, -\frac{x - C_2 t}{C_2}\right) dy \right. \\ &\left. - \frac{1}{h_1} \int_{-h_1}^0 f^{(1)}\left(y, -\frac{x - C_2 t}{C_2}\right) dy \right] \end{aligned} \quad (16b)$$

$$\begin{aligned} A_m^{(1)}(t) &= \frac{-2}{m\pi} \int_{-h_1}^0 \left[ \frac{\partial}{\partial t} f^{(1)}(y, t) \right] \\ &\times \cos\left(\frac{m\pi}{h_1} y\right) dy \end{aligned} \quad (16c)$$

$$\begin{aligned} A_m^{(1)}(t) &= \frac{-2}{m\pi} \int_{-h}^{-h_1} \left[ \frac{\partial}{\partial t} f^{(2)}(y, t) \right] \\ &\times \cos\left(\frac{m\pi}{h_2}(y + h_1)\right) dy \end{aligned} \quad (16d)$$

The functions  $A^{(1)}(x - C_1 t)$  and  $B^{(1)}(x - C_2 t)$  are given in terms of the functions  $A^{(1)}(x - C_1 t)$  and  $B^{(1)}(x - C_2 t)$  by relations (13b) and (13c).

### Free Surface and Interface Elevations of Different Modes

The contributions of the progressive waves of external and internal modes to the free surface elevation,  $\eta_e^{(1)}(x, t)$  and  $\eta_i^{(1)}(x, t)$ , are given using (12) and (14a) as

$$\begin{aligned} \eta_e^{(1)}(x, t) &= \frac{-\varepsilon^2 C_1 \gamma_1}{g(\gamma_1 - \gamma_2)} \\ &\times \frac{\partial}{\partial t} \left[ \frac{\gamma_2}{h_2} \int_{-h}^{-h_1} f^{(2)}\left(y, -\frac{x - C_1 t}{C_1}\right) dy \right. \\ &\left. - \frac{1}{h_1} \int_{-h_1}^0 f^{(1)}\left(y, -\frac{x - C_1 t}{C_1}\right) dy \right], \end{aligned} \quad (17a)$$

$$\begin{aligned} \eta_i^{(1)}(x, t) &= \frac{-\varepsilon^2 C_2 \gamma_2}{g(\gamma_2 - \gamma_1)} \\ &\times \frac{\partial}{\partial t} \left[ \frac{\gamma_1}{h_2} \int_{-h}^{-h_1} f^{(2)}\left(y, -\frac{x - C_2 t}{C_2}\right) dy \right. \\ &\left. - \frac{1}{h_1} \int_{-h_1}^0 f^{(1)}\left(y, -\frac{x - C_2 t}{C_2}\right) dy \right]. \end{aligned} \quad (17b)$$

The contributions of external and internal modes to the interface elevation,  $\eta_e^{(2)}(x, t)$  and  $\eta_i^{(2)}(x, t)$ , are given in

terms of  $\eta_e^{(1)}(x, t)$  and  $\eta_i^{(1)}(x, t)$  using (12)–(14) as

$$\eta_e^{(2)}(x, t) = \frac{\rho_2 - \gamma_1 \rho_1}{\gamma_1 (\rho_2 - \rho_1)} \eta_e^{(1)}(x, t), \quad (18a)$$

$$\eta_i^{(2)}(x, t) = \frac{\rho_2 - \gamma_2 \rho_1}{\gamma_2 (\rho_2 - \rho_1)} \eta_i^{(1)}(x, t). \quad (18b)$$

It is well known, in the studies dealing with stratified fluids, that the external mode is dominant in the neighborhood of the free surface and has a negligible contribution on the interface, while the internal mode is dominant in the neighborhood of the interface and has a negligible contribution on the free surface. If the motion of the wave maker is such that the function  $\eta_e^{(2)}(x, t)$  given by (17a) vanishes, the major contribution of both modes is localized in the neighborhood of the interface and in such cases we say that the wave maker generates internal waves only in the channel.

### Secular Term

A suitable form of the double series representation (11) was used in studying certain problems in the case of homogeneous fluids (Abou-Dina and Helal [3], Abou-Dina and Hassan [1]). It has been shown that this procedure leads to a secular term of the third order in  $\varepsilon$  in the expression of the velocity potential. This secular term, which increases indefinitely with the increase of the variable  $x$ , vanishes at  $x = 0$ . The same result can be shown in the case of stratified fluids, according to the analysis of Abou-Dina and Helal [2]. Hence, although expressions (12)–(18) are valid at  $x = 0$ , they are not adequate for the description of the propagation of waves far from the wave maker. This unacceptable result is due to certain aspects of the mathematical procedure used (see Abou-Dina and Helal [3], Abou-Dina and Hassan [1] for the justification) and needs to be remedied.

Our main interest, in the remaining part of the present paper, is to modify the mathematical procedure used above in order to describe the propagation of waves generated by the wave maker in going towards the down-stream extremity of the channel.

### Multiple Scale Transformation of Variables

We use the set of variables  $u, v$  and  $y$  defined in terms of the distorted set  $x, y, t$  by (cf. Benney [8] and Temperville [34]):

$$u = x - Ct, \quad v = \varepsilon^2 x, \quad (19)$$

where  $C$  is a real constant to be precised.

Equations (2), (4) up to (8) can be written in terms of  $u, v$  and  $y$  as:

(i) In the fluid mass:

$$\varepsilon^6 \partial_{vv} \Phi^{(j)} + 2\varepsilon^4 \partial_{uv} \Phi^{(j)} + \varepsilon^2 \partial_{uu} \Phi^{(j)} + \partial_{yy} \Phi^{(j)} = 0. \quad (20)$$

(ii) On the free surface:

$$\begin{aligned} \partial_y \Phi^{(1)} &= \varepsilon^6 \partial_v \eta^{(1)} \partial_v \Phi^{(1)} \\ &\quad + \varepsilon^4 \left\{ \partial_u \eta^{(1)} \partial_v \Phi^{(1)} + \partial_v \eta^{(1)} \partial_u \Phi^{(1)} \right\} \\ &\quad + \varepsilon^2 \partial_u \eta^{(1)} \partial_u \Phi^{(1)} - \varepsilon C \partial_u \eta^{(1)} \\ &\quad \text{at } y = \eta^{(1)}(u, v), \end{aligned} \quad (21a)$$

$$\begin{aligned} g \eta^{(1)} - \varepsilon C \partial_u \Phi^{(1)} + \frac{1}{2} \left[ \varepsilon^2 \left\{ \partial_u \Phi^{(1)} + \varepsilon^2 \partial_v \Phi^{(1)} \right\}^2 \right. \\ \left. + (\partial_y \Phi^{(1)})^2 \right] = 0 \quad \text{at } y = \eta^{(1)}(u, v). \end{aligned} \quad (21b)$$

(iii) On the interface:

$$\begin{aligned} \partial_y \Phi^{(j)} &= \varepsilon^6 \partial_v \eta^{(2)} \partial_v \Phi^{(j)} \\ &\quad + \varepsilon^4 \left\{ \partial_u \eta^{(2)} \partial_v \Phi^{(j)} + \partial_v \eta^{(2)} \partial_u \Phi^{(j)} \right\} \\ &\quad + \varepsilon^2 \partial_u \eta^{(2)} \partial_u \Phi^{(j)} - \varepsilon C \partial_u \eta^{(2)} \\ &\quad \text{at } y = \eta^{(1)}(u, v) \text{ for } j = 1, 2 \end{aligned} \quad (22a)$$

$$\begin{aligned} \rho_1 \left\{ g \left( h_1 + \eta^{(2)} \right) - \varepsilon C \partial_u \Phi^{(1)} \right. \\ \left. + \frac{1}{2} \left[ \varepsilon^2 \left\{ \partial_u \Phi^{(1)} + \varepsilon^2 \partial_v \Phi^{(1)} \right\}^2 + (\partial_y \Phi^{(1)})^2 \right] \right\} \\ = \rho_2 \left\{ g \left( h_1 + \eta^{(2)} \right) - \varepsilon C \partial_u \Phi^{(2)} \right. \\ \left. + \frac{1}{2} \left[ \varepsilon^2 \left\{ \partial_u \Phi^{(2)} + \varepsilon^2 \partial_v \Phi^{(2)} \right\}^2 + (\partial_y \Phi^{(2)})^2 \right] \right\} \\ \text{at } y = -h_1 + \eta^{(2)}(u, v). \end{aligned} \quad (22b)$$

(iv) On the bottom:

$$\partial_y \Phi^{(2)} = 0 \quad \text{at } y = -h. \quad (23)$$

In the regions where we are interested in far from the wave maker, the contribution of the local disturbance on the motion of the fluid layers can be neglected and we use the



following representations for the functions  $\Phi^{(j)}(u, v, y)$  and  $\eta^{(j)}(u, v)$  (cf. Temperville [34]):

$$\Phi^{(j)}(u, v, y) = \sum_{n=1}^{\infty} \varepsilon^{2n-1} \Phi_{2n-1}^{(j)}(u, v, y), \quad (24a)$$

$$\eta^{(j)}(u, v) = \sum_{n=1}^{\infty} \varepsilon^{2n} \eta_{2n}^{(j)}(u, v). \quad (24b)$$

### Derivation of the KdV Equation

It can be shown from the above equations that for  $n = 1$ , the functions  $\Phi_1^{(j)}(u, v, y)$ ,  $j = 1, 2$  are independent of the variable  $y$  and that

$$\eta_2^{(1)}(u, v) = \frac{C}{g} \partial_u \Phi_1^{(1)}(u, v, y), \quad \text{at } y = 0, \quad (25)$$

$$\eta_2^{(2)}(u, v) = \frac{C}{g(\rho_2 - \rho_1)} \left[ \rho_2 \partial_u \Phi_1^{(2)}(u, v, y) - \rho_1 \partial_u \Phi_1^{(1)}(u, v, y) \right], \quad \text{at } y = -h_1. \quad (26)$$

Also, for  $n = 2$ , the values of the constant  $C$  appearing in (19) can be obtained as

$$C = \begin{cases} C_1 & \text{for the barotropic (external) mode,} \\ C_2 & \text{for the baroclinic (internal) mode,} \end{cases} \quad (27)$$

where  $C_j$  is given by (13a) for  $j = 1, 2$ .

It can also be shown that the results up to  $n = 3$  contain no secular terms, and that the function  $\eta_2^{(1)}(u, v)$ , characterizing the free surface elevation, satisfies the following KdV equation:

$$L \partial_{uuu} \eta_2^{(1)} + M \eta_2^{(1)} \partial_u \eta_2^{(1)} + N \partial_v \eta_2^{(1)} = 0, \quad (28)$$

where

$$L = \frac{1}{6} (2h^3 + h_1^3 - 3hh_1^2) + \frac{\gamma}{2} \left( \frac{C^2}{g} - \frac{h_1}{3} \right), \quad (29a)$$

$$M = 3\gamma + \frac{2(1-\gamma)(\rho_2 - \gamma\rho_1)}{\gamma(\rho_2 - \rho_1)}, \quad (29b)$$

$$N = 2(h_2 + \gamma h_1) \quad (29c)$$

and

$$\gamma = \frac{g h_2}{C^2 - g h_1}. \quad (29d)$$

The initial condition at  $v = 0$  (or  $x = 0$  and  $u = -Ct$ ), needed for the solution of the KdV Eq. (28) is obtained from (17) in the form:

- Barotropic mode:

$$\eta_e^{(1)}(u, 0) = \frac{\varepsilon^2 C_1^2 \gamma_1}{g(\gamma_1 - \gamma_2)} \times \frac{\partial}{\partial u} \left[ \frac{\gamma_2}{h_2} \int_{-h}^{-h_1} f^{(2)} \left( y, -\frac{u}{C_1} \right) dy - \frac{1}{h_1} \int_{-h_1}^0 f^{(1)} \left( y, -\frac{u}{C_1} \right) dy \right] \quad (30a)$$

- Baroclinic mode:

$$\eta_i^{(1)}(u, 0) = \frac{\varepsilon^2 C_2^2 \gamma_2}{g(\gamma_2 - \gamma_1)} \times \frac{\partial}{\partial u} \left[ \frac{\gamma_1}{h_2} \int_{-h}^{-h_1} f^{(2)} \left( y, -\frac{u}{C_2} \right) dy - \frac{1}{h_1} \int_{-h_1}^0 f^{(1)} \left( y, -\frac{u}{C_2} \right) dy \right] \quad (30b)$$

The following relation is also obtained between  $\eta_2^{(2)}(u, v)$  and  $\eta_2^{(1)}(u, v)$ :

$$\eta_2^{(2)}(u, v) = \frac{\rho_2 - \gamma\rho_1}{\gamma(\rho_2 - \rho_1)} \eta_2^{(1)}(u, v). \quad (31)$$

### Conclusions

A multiple scale transformation of variables and a single series representations for the velocity potentials and the free surface and the interface elevations show that these elevations up to the second order satisfy the KdV equation. The initial conditions needed for the solution of this equation are obtained at the position of the wave maker using another technique depending on a double series representation of the unknown functions. The particular choice of the motion of the wave maker can minimize the elevation of the free surface and hence lead to dominant nonlinear internal waves only in the channel.

### Future Directions

For a future work following the present one, it is intended to investigate the following items:

- Study the possibility of generating nonlinear internal waves by different types of the motion of the plane wave maker. This will need analytical and numerical work as well.
- Study the inverse problem, precisely the possibility of the recuperation of the water waves' energy and employing it in producing a controlled solid body motion.

- Study the possibility of generating nonlinear internal waves in model problems with different and more complicated geometry, to simulate the actual physical situations.

## Bibliography

### Primary Literature

1. Abou-Dina MS, Hassan FM (2006) Generation and propagation of nonlinear tsunamis in shallow water by a moving topography. *Appl Math Comput* 177:785–806
2. Abou-Dina MS, Helal MA (1990) The influence of submerged obstacle on an incident wave in stratified shallow water. *Eur J Mech B/Fluids* 9(6):545–564
3. Abou-Dina MS, Helal MA (1992) The effect of a fixed barrier on an incident progressive wave in shallow water. *Il Nuovo Cimento* 107B(3):331–344
4. Abou-Dina MS, Helal MA (1995) The effect of a fixed submerged obstacle on an incident wave in stratified shallow water (Mathematical Aspects). *Il Nuovo Cimento B* 110(8): 927–942
5. Barthelemy E, Kabbaj A, Germain JP (2000) Long surface wave scattered by a step in a two-layer fluid. *Fluid Dyn Res* 26: 235–255
6. Benjamin TB (1966) Internal waves of finite amplitude and permanent form. *J Fluid Mech* 25:241–270
7. Benjamin TB (1967) Internal waves of permanent form of great depth. *J Fluid Mech* 29:559–592
8. Benney DJ, LIN CC (1960) On the secondary motion induced by oscillations in a shear flow. *Phys Fluids* 3:656–657
9. Boussinesq MJ (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire. *Acad Sci Paris, CR Acad Sci* 72:755–759
10. Chen Y, Liu PL-F (1995) Modified Boussinesq equations and associated parabolic models for water wave propagation. *J Fluid Mech* 288:351–381
11. Choi W, Camassa R (1996) Weakly nonlinear internal waves in a two-fluid system. *J Fluid Mech* 313:83–103
12. Choi W, Camassa R (1999) Fully nonlinear internal waves in a two-fluid system. *J Fluid Mech* 396:1–36
13. Craig W, Guyenne P, Kalisch H (2005) Hamiltonian long-wave expansions for free surfaces and interfaces. *Commun Pure Appl Math* 18:1587–1641
14. Davis RE, Acrivos A (1967) Solitary internal waves in deep water. *J Fluid Mech* 29:593–607
15. Garrett C, Munk W (1979) Internal waves in the ocean. *Ann Rev Fluid Mech* 11:339–369
16. Germain JP (1971) Sur le caractère limite de la théorie des mouvements des liquides parfaits en eau peu profonde. *CR Acad Sci Paris Série A* 273:1171–1174
17. Germain JP (1972) Théorie générale d'un fluide parfait pesant en eau peu profonde de profondeur constante. *CR Acad Sci Paris Série A* 274:997–1000
18. Gobbi MF, Kirby JT, Wei G (2000) A fully nonlinear Boussinesq model for surface waves-Part 2. Extension to  $O(kh)^4$ . *J Fluid Mech* 405:181–210
19. Helal MA, Moline JM (1981) Nonlinear internal waves in shallow water: A theoretical and experimental study. *Tellus* 33:488–504
20. Kabbaj A (1985) Contribution à l'étude du passage des ondes des gravité sur le talus continental et à la generation des ondes internes. These de doctorat d'état, IMG Université de Grenoble
21. Keulegan GH (1953) Hydrodynamical effects of gales on Lake Erie. *J Res Natl Bur Std* 50:99–109
22. Kubota T, Ko DRS, Dobbs LD (1978) Propagation of weakly nonlinear internal waves in a stratified fluid of finite depth. *AIAA J Hydrodyn* 12:157–165
23. LeBlond PH, Mysak LA (1978) *Waves in Ocean*. Elsevier, Amsterdam
24. Liu C-M, Lin M-C, Kong C-H (2008) Essential properties of Boussinesq equations for internal and surface waves in a two-fluid system. *Ocean Eng* 35:230–246
25. Long RR (1956) Solitary waves in one- and two-fluid systems. *Tellus* 8:460–471
26. Lynett PJ, Liu PL-F (2002) A two-dimensional depth-integrated model for internal wave propagation over variable bathymetry. *Wave Motion* 36:221–240
27. Madsen PA, Schaffer HA (1998) Higher-order Boussinesq-type equations for surface gravity waves: derivation and analysis. *Philos Trans R Soc Lond A* 356:3123–3184
28. Matsuno Y (1993) A unified theory of nonlinear wave propagation in two-layer fluid systems. *J Phys Soc Jpn* 62:1902–1916
29. Nwogu O (1993) Alternative form of Boussinesq equations for nearshore wave propagation. *J Waterways Port Coast Ocean Eng ASCE* 119:618–638
30. Ono H (1975) Algebraic solitary waves in stratified fluids. *J Phys Soc Jpn* 39:1082–1091
31. Peters AS, Stoker JJ (1960) Solitary waves in liquid having non-constant density. *Comm Pure Appl Math* 13:115–164
32. Robinson RM (1969) The effect of a vertical barrier on internal waves. *Deep-Sea Res* 16:421–429
33. Stoker JJ (1957) *Water waves*. Interscience, New York
34. Temperville A (1985) Contribution à la théorie des ondes de gravité en eau peu profonde. Thèse de doctorat d'état, IMG Université de Grenoble
35. Wehausen JV, Laitone EV (1960) Surface waves. In: *Handbuch der Physik* 9. Springer, Berlin
36. Wei G, Kirby JT, Grilli ST, Subramanya R (1995) A fully nonlinear Boussinesq model for surface waves, Part 1. Highly nonlinear, unsteady waves. *J Fluid Mech* 294:71–92

### Books and Reviews

- Germain JP, Guli L (1977) Passage d'une onde sur une barrier mince immergée en eau peu profonde. *Ann Hydrog* 5(746):7–11
- Pinettes M-J, Renouard D, Germain J-P (1995) Analytical and experimental study of the oblique passing of a solitary wave over a shelf in a two-layer fluid. *Fluid Dyn Res* 16:217–235

## Non-linear Ordinary Differential Equations and Dynamical Systems, Introduction to

FERDINAND VERHULST

Mathematisch Instituut, University of Utrecht,  
Utrecht, The Netherlands

An ordinary differential equation (ODE) is called linear if it can be written in the form  $dy/dx = f(x)y + g(x)$  with  $x$  a real or complex variable and  $y$  an  $n$ -dimensional (finite) real or complex vector function. Non-linear ODEs are then  $n$ -dimensional ODEs of the form  $dy/dx = F(x, y)$  that are not linear. Jean Mawhin famously compared this distinction to a division of the animal world into ‘elephants’ and ‘non-elephants’, but even admitting that the distinction between linear and non-linear ODEs is artificial and of relatively recent date, it makes a little bit more sense than it looks like at first sight. The reason is, that in many problem formulations in classical physics, linear equations are quite common. Also, when considering non-linear ODEs, but linearizing around particular solutions, a number of fundamental theorems can be used to characterize the particular solutions starting with the features of the linearized equation. This will become clear in a number of the articles that follow. On the other hand, right from the beginning of the development of classical physics, the analysis of differential equations with their fully non-linear behavior has been necessary and essential. Think of celestial mechanics that started in the 18th century and for instance the analysis of solitons in fluid mechanics.

The great scientists of the 18th century, among which Newton, Euler, Lagrange and Laplace, were all concerned with the formulation and first analysis steps of non-linear ODEs. This work was continued by mathematicians like Jacobi and Painlevé who devoted much of their attention to transformation methods and the analysis of special cases. A new stimulus came in the second half of the 19th century with the insights and fundamental ideas of Henri Poincaré. In fact his ideas are still fully alive and active today. Poincaré realized that non-linear ODEs can in general not be solved explicitly, i. e. expressed in terms of elementary functions, so that calculations should be supplemented by qualitative theory. Poincaré developed both quantitative and qualitative methods and his approach has shaped the analysis of non-linear ODEs in the period that followed up till now, becoming part of the topic of dynamical systems. This volume of the Encyclopedia reflects to a large part these new developments.

Basic questions on existence and uniqueness of solutions are discussed by Gianne Derks (see ► [Existence and Uniqueness of Solutions of Initial Value Problems](#)) with attention to recent extensions of the notion of an ODE. Most of this is classical material.

The article by Carmen Chicone (see ► [Stability Theory of Ordinary Differential Equations](#)) deals with stability theory. It is concerned with the mathematical formulation and the basic results, but also includes the stability ques-

tion in the context of conservative systems and the part played by the KAM theorem. The stability of periodic orbits gets special attention and in addition the stability of the orbit structure as a whole. This is usually referred to as structural stability.

The theory of periodic solutions of non-autonomous ODEs displays a number of striking differences with the autonomous case. Jean Mawhin (see ► [Periodic Solutions of Non-autonomous Ordinary Differential Equations](#)) presents classical and new results in this field.

A more general technique to study equilibria, periodic solutions and their bifurcations is the Lyapunov–Schmidt method which is intimately connected to conditions of the implicit function theorem in its abstract form. André Vanderbauwhede (see ► [Lyapunov–Schmidt Method for Dynamical Systems](#)) explains this basic method for equilibria and for periodic solutions with extensions to infinite dimensions.

Center manifolds are manifolds associated with the critical part of the spectrum of equilibria and periodic solutions. They arise naturally in theory and applications. George Osipenko (see ► [Center Manifolds](#)) discusses the theory with special attention to the corresponding (and very effective) reduction principle.

A number of special topics have been getting a lot of attention, both in mathematical and in applied research. Relaxation oscillations are described by Johan Grasman (see ► [Relaxation Oscillations](#)), starting with the classical example of the Van der Pol-oscillator and going on to more complicated systems. This also involves the discussion of canards in geometric singular perturbation theory which continues to produce activity in dynamical systems research and applications in mathematical biology. The dynamics of Hamiltonian systems is described by Heinz Hanßmann (see ► [Dynamics of Hamiltonian Systems](#)). Starting with the classical formulation and basic notions, the article elaborates on integrability questions and the KAM theorem, Cantor dust and tori bifurcations, thus introducing the reader to the most recent results.

Periodic solutions of Hamiltonian systems merit a special treatment in the article by Luca Sbano (see ► [Periodic Orbits of Hamiltonian Systems](#)). Natural problems are the continuation of orbits and the part played by symmetries. Variational methods have become important in this field, enabling us to consider the topology and geometry of periodic paths in sufficient generality.

Another classical topic that is fully alive is the dynamics of parametric excitation. Alan Champneys (see ► [Dynamics of Parametric Excitation](#)) introduces the necessary Floquet theory, corresponding bifurcation diagrams and a number of applications.

Non-linear ODEs has grown into the more general field of dynamical systems. Beginning with hyperbolic behavior and examples, Araújo and Viana (see ► [Hyperbolic Dynamical Systems](#)) discuss attractors and physical measures with the idea of stochastic stability. This leads naturally to the formulation of obstructions to hyperbolicity, partial and non-uniform hyperbolicity.

To handle dynamical systems in practice, one needs quantitative methods. Meijer, Dercole and Oldeman (see ► [Numerical Bifurcation Analysis](#)) consider numerical bifurcation analysis for systems depending on parameters. This involves continuation and detection of bifurcations, complications like branch switching and orbit connection and a guide to possible software environments. The literature on non-linear ODEs grows at a fast pace. The articles are presenting a tour of the relevant literature, and even more importantly, they point out future directions of research in their respective fields.

## Non-linear Partial Differential Equations, Introduction to

ITALO CAPUZZO DOLCETTA  
Dipartimento di Matematica,  
Sapienza Università di Roma, Rome, Italy

A large number of nonlinear phenomena in fundamental sciences (physics, chemistry, biology ...), in technology (material science, control of nonlinear systems, ship and aircraft design, combustion, image processing ...) as well in economics, finance and social sciences are conveniently modeled by nonlinear partial differential equations (NLPDE, in short). Let us mention, among the most important examples for the applications and from the historical point of view, the Euler and Navier–Stokes equations in fluid dynamics and the Boltzmann equation in gas dynamics. Other fundamental models, just to mention a few of them, are reaction-diffusion, porous media, nonlinear Schrödinger, Klein–Gordon, eikonal, Burger and conservation laws, nonlinear wave Korteweg–de Vries ...

The above list is by far incomplete as one can easily realize by looking at the current scientific production in the field as documented, for example, by the American Mathematical Society database **MathSciNet**.

Despite an intense mathematical research activity going on since the second half of the XXth century, the extremely diversified landscape of the area of NLPDE is still largely unexplored and a number of difficult problems are still open.

Generally (and therefore rather vaguely) speaking, one of the main difficulties induced by the nonlinear (as opposed to linear) structure of a partial differential equation is that such equations, especially first-order ones, do not in general possess smooth solutions. The analytical approach therefore requires the adoption of appropriate generalized notions of solutions (weak solutions in Sobolev spaces, entropy solutions, viscosity solutions, re-normalized solutions ...) in order to handle the classical Hadamard well-posedness approach to the rigorous validation of the model (existence, uniqueness and stability of solutions).

Another fundamental difference with respect to the linear theory lies in the fact that no explicit representation formulas for solutions are in general available, even for simplified models. Fortunately enough however, **good theory is (almost) as useful as exact formulas**, as stated by L.C. Evans in the preface of his book *Partial Differential Equations, Graduate Studies in Mathematics* (Volume 19, American Mathematical Society, Providence Rhode Island, 1998). It is clear also, in this respect, that progress in NLPDE is necessarily strongly connected with the development of numerical methods of simulations.

A further “pathology” due to nonlinearity is that solutions of initial value problems for nonlinear evolution equations may exist only for a small lap of time (the blow-up phenomenon). These and other unavoidable features imply the need to develop new, sophisticated and often ad hoc mathematical methods for investigating different classes of NLPDE.

The nonlinear theory (rather, theories) has been widely developed for different classes of equations, mainly in the direction of existence (via monotonicity, compactness, critical points methods ...), regularity of generalized solutions, qualitative analysis of solutions, limit problems (e. g. large time asymptotic, homogenization, scaling limits ...).

Quoting P.L. Lions, **new tools are required and discovered [at this purpose] in connection with all fields of analysis (functional, real, global, complex, numerical, Fourier ...)**, see *On Some Challenging Problems in Non-linear Partial Differential Equations* (Mathematics: Frontiers and Perspectives 2000, International Mathematical Union).

The 11 papers comprised in this section offer deep insight and up to date information on some of the various aspects mentioned above. A common feature of the papers is that their focus is on nonlinear partial differential equations arising in “real world” applications. The paper ► [Hyperbolic Conservation Laws](#) by A. Bressan reports on very recent important advances in the understanding of first-order nonlinear partial differential equations (or systems) describing the time evolution of certain basic quantities in

physical models such as mass, momentum, energy, electric charge. The main mathematical difficulty is related to the strong linearity and the absence of dissipative terms in the equations; the evolution of smooth initial data may then produce discontinuities (shocks) in finite time, thus requiring the introduction of appropriate notion of weak entropic solutions.

Many important NLPDE, e.g. Hamilton–Jacobi–Bellman and Isaacs equations arising in optimal control and game theory, mean-curvature flow and  $\infty$ -Laplace equations do not have a divergence structure. In this case, the notion of weak solution in the distribution sense is not applicable. S. Koike’s [► Non-linear Partial Differential Equations, Viscosity Solution Method in](#) provides an account of ideas and results of the theory of viscosity solutions. This generalized notion of solution, which is intimately related to the maximum principle for elliptic equations, has proved to be most appropriate to implement the well-posedness program to large classes of second-order fully NLPDE not treatable by variational methods.

First-order NLPDE of Hamilton–Jacobi-type have a central relevance both in classical and quantum mechanics as well in calculus of variations and optimal control theory. Lack of regularity occurs here at the level of the smoothness of the gradient of solutions for reasons which are somewhat similar to those indicated for hyperbolic conservation laws. The contribution by A. Siconolfi [► Hamilton–Jacobi Equations and Weak KAM Theory](#) reports on current research on the impact of the notion of the Aubry–Mather set, imported from dynamical systems theory, in the understanding of uniqueness, extra smoothness properties and large time behavior of viscosity solutions.

In a physical model, dispersion can be understood as the decrease with time of the size of some relevant quantity such as matter or energy over a given volume. P. D’Ancona’s [► Dispersion Phenomena in Partial Differential Equations](#) examines in detail analytical techniques to evaluate, in suitable norms, the rate of dispersion in some very fundamental physical models such as the nonlinear wave and Schrödinger equations; new tools in Fourier and harmonic analysis play a key role in this respect. In connection with a general remark made above, it is worth pointing out that sharp dispersion (or Strichartz type) estimates have relevant applications to the issue of existence of solutions that are global in time.

The interaction of nonlinearity with randomness in a PDE model is the leading theme of [► Non-linear Stochastic Partial Differential Equations](#) by G. Da Prato. This is a definitely new and rapidly developing area of investigation. The paper presents a general nonlinear semi-

group approach to the study, mainly focused at global existence and well-posedness, of a large class of stochastic dynamical systems on Hilbert spaces. The large time behavior of solutions is also analyzed, relying on the notion of invariant measure. The application of general abstract results to specific models of interest, such as Ornstein–Uhlenbeck, reaction-diffusion, Burger’s, 2D Navier–Stokes equations perturbed by noise is also discussed in the paper.

The Navier–Stokes equations of fluid dynamics are presently a fundamental model for simulations in several branches of applied sciences (e.g. meteorology, oceanography, biology ...) and design in industry (airplane, car, oil ...). The text by G.P. Galdi [► Navier–Stokes Equations: A Mathematical Analysis](#) is an extended and fairly complete review of mathematical tools (appropriate function spaces, a priori estimates, fixed point theorems), results (concerning in particular the well-posedness of initial and boundary value problems) and open questions (including the famous one concerning the global regularity of solutions in 3D) pertaining to this important topic.

Stokes equations, which are obtained as a formal limit from Navier–Stokes equations, are central in the article [► Biological Fluid Dynamics, Non-linear Partial Differential Equations](#) by A. De Simone, F. Alouges, and A. Lefebvre. A mathematical model for the swimming of a micro-organism at a low Reynolds number regime is proposed. Methods from sub-Riemannian geometry as well as numerical ones for the quantitative optimization of the strokes of micro-swimmers are discussed.

Control theory is concerned with the issue of influencing, by way of an external action, the evolution of a system. Many applications, both from the scientific and technological point of view, involve distributed control systems, that is systems whose evolution is governed by partial differential equations. Important examples range from control of diffusion processes to nonlinear elasticity, from control of biological systems (see above) to traffic flows on networks. The text by Alabau-Boussouira and P. Cannarsa [► Control of Non-linear Partial Differential Equations](#) is a wide-ranging overview of the state-of-the-art in the field, covering among others the issues of controllability, stabilization and optimal control.

Traffic flows on networks are discussed in [► Vehicular Traffic: A Review of Continuum Mathematical Models](#) by B. Piccoli and A. Tosin. The paper reviews in particular macroscopic and kinetic models. Macroscopic modeling of traffic flows stems from the Euler and Navier–Stokes equations while kinetic models rely on principles of statistical mechanics. Since macroscopic models take often the form of nonlinear hyperbolic conservation laws, special at-



tention is given in the paper to the analysis of those models by nonlinear hyperbolic equation techniques (see also the paper by Bressan in this respect).

An important issue in modeling nonlinear physical systems is the reduction of complexity. One way to realize this is to use scaling limit procedures allowing a more manageable description of the system itself through macroscopic coordinates. The paper [Scaling Limits of Large Systems of Non-linear Partial Differential Equations](#) contributed by D. Benedetto and M. Pulvirenti illustrates ideas and methods to treat large classical and quantum particle systems in the weak coupling regime. It is also shown that those systems can be conveniently described by macroscopic quantities whose evolution is governed by the Fokker–Planck–Landau and the Boltzmann equations for the classical and quantum case, respectively.

## Non-linear Partial Differential Equations, Viscosity Solution Method in

SHIGEAKI KOIKE

Department of Mathematics, Saitama University,  
Saitama, Japan

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Examples  
Comparison Principle  
Existence Results  
Boundary Value Problems  
Asymptotic Analysis  
Other Notions  
Future Directions  
Bibliography

### Glossary

**Weak solutions** In the study of  $m$ th order partial differential equations (abbreviated, PDEs), a function is informally called a classical solution of the PDE if (a) it is  $m$ -times differentiable, and (b) the PDE holds at each point of the domain by putting its derivatives there. However, it is not easy to find such classical solutions of PDEs in general except for some special cases. The standard strategy to find a classical solution is first to look for a candidate of solutions, which becomes

a classical solution if the property (a) holds for it. Here, such a candidate is called a weak solution of the PDE.

**Viscosity solutions** In 1981, for first order PDEs of non-divergence type, M.G. Crandall and P.-L. Lions in [\[15,16\]](#) (also [\[17\]](#)) introduced the notion of weak solutions, which are called viscosity solutions. The definition of those is the property which the limit of approximate solutions via the vanishing viscosity method admits.

Afterwards, the notion was extended to fully nonlinear second order elliptic/parabolic PDEs.

For general theory of viscosity solutions, [\[3,5,7,18,20\]](#) are recommended for the interested readers.

Throughout this article, to minimize the references, [\[18\]](#) will be often referred to instead of the original papers except for some pioneering works or those which appeared after [\[18\]](#).

**Ellipticity/parabolicity** General second order PDEs under consideration are

$$F(x, u(x), Du(x), D^2u(x)) = 0 \quad \text{in } \Omega. \quad (\text{E})$$

Here,  $u: \Omega \rightarrow \mathbf{R}$  is the unknown function,  $\Omega \subset \mathbf{R}^n$  an open set,  $F: \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \rightarrow \mathbf{R}$  a given (continuous) function,  $S^n$  the set of  $n \times n$  symmetric matrices with the standard ordering,  $Du(x) = ((\partial u)/(\partial x_1)(x), \dots, (\partial u)/(\partial x_n)(x))$ , and  $D^2u(x) \in S^n$  whose  $(i, j)$ th entry is  $(\partial^2 u)/(\partial x_i \partial x_j)(x)$ .

According to the early literature in viscosity solution theory, (E) is called elliptic if

$$X \leq Y \implies F(x, r, p, X) \geq F(x, r, p, Y)$$

for  $(x, r, p, X, Y) \in \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \times S^n$ . It should be remarked that the opposite order has been also used.

When  $F$  does not depend on the last variables (i.e. first order PDEs), it is automatically elliptic. Thus, the above notion has been called degenerate elliptic.

The evolution version of general PDEs is

$$u_t(x, t) + F(x, t, u(x, t), Du(x, t), D^2u(x, t)) = 0 \quad \text{in } Q_T := \Omega \times (0, T]. \quad (\text{P})$$

Here,  $u: Q_T \rightarrow \mathbf{R}$  is the unknown function,  $F: \Omega \times (0, T] \times \mathbf{R} \times \mathbf{R}^n \times S^n \rightarrow \mathbf{R}$  a given function,  $T > 0$ , and  $u_t(x, t) = (\partial u)/(\partial t)(x, t)$ . In (P), the notations  $Du(x, t)$  and  $D^2u(x, t)$  do not contain derivatives with respect to  $t$ .

Similarly, (P) is called parabolic if  $F(\cdot, t, \cdot, \cdot, \cdot)$  is elliptic for each  $t \in (0, T]$ .

**Uniform ellipticity/uniform parabolicity** Denoted by  $S_{\lambda, \Lambda}^n := \{A \in S^n \mid \lambda I \leq A \leq \Lambda I\}$  for fixed  $0 < \lambda \leq \Lambda$ , the Pucci operators  $\mathcal{P}^\pm: S^n \rightarrow \mathbf{R}$  are defined by

$$\mathcal{P}^+(X) := \max_{A \in S_{\lambda, \Lambda}^n} \{-\text{trace}(AX)\} \quad \text{and}$$

$$\mathcal{P}^-(X) := \min_{A \in S_{\lambda, \Lambda}^n} \{-\text{trace}(AX)\}.$$

Then, the PDE (E) is called uniformly elliptic if

$$\mathcal{P}^-(X - Y) \leq F(x, r, p, X) - F(x, r, p, Y) \leq \mathcal{P}^+(X - Y)$$

for  $(x, r, p, X, Y) \in \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \times S^n$ . This is a fully nonlinear version of the standard uniform ellipticity. For the theory of second order uniformly elliptic PDEs, [24] is the standard text book.

Similarly, (P) is called uniformly parabolic if  $F(\cdot, t, \cdot, \cdot, \cdot)$  is uniformly elliptic for each  $t \in (0, T]$ .

**Dynamic programming principle** In stochastic control problems, the value function is determined by minimizing given cost functionals. The dynamic programming principle (abbreviated, DPP), which was established as the Bellman's principle of optimality, is a formula which the value function satisfies.

The DPP indicates that the value function is a viscosity solution of the associated Hamilton–Jacobi–Bellman (abbreviated, HJB) equation.

## Definition of the Subject

In order to investigate phenomena in both natural and social sciences, it is important to analyze solutions of PDEs derived from certain minimization principles such as calculus of variations.

Since it is hard to find classical solutions of PDEs in general, the first strategy is to look for weak solutions of those. For PDEs of divergence type, the most celebrated notion of weak solutions is that in the distribution sense, which is formally derived through the integration by parts.

On the other hand, before viscosity solutions were introduced, there were several notions of weak solutions for PDEs of non-divergence type such as generalized solutions for first order PDEs by S.N. Kružkov.

For second order elliptic/parabolic PDEs of non-divergence type, it has turned out through much research that the notion of viscosity solutions is the most appropriate one by several reasons:

- (i) Viscosity solutions admit well-posedness (i. e. existence, uniqueness, stability),
- (ii) The notion of viscosity solutions is naturally derived from the vanishing viscosity method,

- (iii) In deterministic/stochastic control problems, the expected solutions are the unique viscosity solutions of the associated HJB equations in principle,
- (iv) The viscosity solution method can deal with first and second order elliptic/parabolic PDEs by a unified manner,
- (v) The viscosity solution method can be applied to various non-divergent PDEs for which there were no good notions of weak solutions such as HJB equations, mean curvature flow equations,  $\infty$ -Laplace equation etc.

## Introduction

In what follows, elliptic PDEs (E) will be mainly discussed since one can treat (P) with some modifications. For the sake of simplicity, the variables of the unknown functions and their derivatives in the PDEs will be suppressed if there is no confusion.

For  $U \subset \mathbf{R}^n$ , which will be  $\Omega$ ,  $\overline{\Omega}$  or  $\partial\Omega$  later,  $C^1(U)$  (resp.,  $C^2(U)$ ) denotes the set of real-valued functions  $\phi$  in  $U$  such that  $\phi$  and  $D\phi$  (resp.,  $\phi$ ,  $D\phi$  and  $D^2\phi$ ) are continuous in  $U$ .

The definition of viscosity solutions presented in [17] has been widely used although the original one in [15,16] is different but equivalent.

**Definition 1**  $u: \Omega \rightarrow \mathbf{R}$  is called a viscosity subsolution (resp., supersolution) of (E) in  $\Omega$  if

$$F(x, u(x), D\phi(x), D^2\phi(x)) \leq 0 \quad (\text{resp., } \geq 0) \quad (1)$$

whenever  $(u - \phi)(x) = \sup_{\Omega} (u - \phi)$  (resp.,  $\inf_{\Omega} (u - \phi)$ ) for  $\phi \in C^2(\Omega)$  and  $x \in \Omega$ .

Finally,  $u$  is called a viscosity solution of (E) in  $\Omega$  if it is a viscosity sub- and supersolution of (E) in  $\Omega$ .

**Remark 2** In order to study boundary value problems in the viscosity sense in Sect. “Boundary Value Problems”, it is necessary to modify the above definition by replacing  $\Omega$  by  $\overline{\Omega}$ . To this end,  $\Omega$  in the semi-jets, the equivalent definition and Ishii's lemma etc. below should be also replaced by  $\overline{\Omega}$ .

The above definition indicates that if the so-called test function  $\phi \in C^2(\Omega)$  touches  $u$  from above (resp., below) at  $x \in \Omega$ , by noting  $0 = (u - \phi)(x) \geq$  (resp.,  $\leq$ )  $(u - \phi)(y)$  for  $y \in \Omega$ , the one-sided inequality (1) is required, instead of the Eq. (E), with the derivatives of  $\phi$  at  $x \in \Omega$ .

It is worth mentioning that if  $u \in C^2(\Omega)$  is a viscosity subsolution (resp., supersolution) of (E) in  $\Omega$ , then it is

a classical subsolution (resp., supersolution) of (E) in  $\Omega$ ;

$$F(x, u(x), Du(x), D^2u(x)) \leq 0 \quad (\text{resp., } \geq 0) \\ \text{for each } x \in \Omega.$$

On the contrary, if  $u \in C^2(\Omega)$  is a classical subsolution (resp., supersolution) of (E) in  $\Omega$ , then it is a viscosity subsolution (resp., supersolution) of (E) in  $\Omega$  whenever (E) is elliptic.

One can observe that if  $u$  is a viscosity subsolution (resp., supersolution) of (E) in  $\Omega$ , then  $-u$  is a viscosity supersolution (resp., subsolution) of

$$-F(x, -u, -Du, -D^2u) = 0 \quad \text{in } \Omega.$$

If  $F$  is independent of  $X$  (i.e. first order PDEs), then it is possible to replace  $C^2(\Omega)$  by  $C^1(\Omega)$  in the definition.

For the definition to (P),  $\Omega$  and  $C^2(\Omega)$ , respectively, are changed by  $Q_T$  and  $C^{2,1}(Q_T)$  in the above, where  $Q_T = \Omega \times (0, T]$ . Here,  $C^{2,1}(Q_T)$  denotes the set of real-valued functions  $\phi$  in  $Q_T$  such that  $\phi$ ,  $D\phi$ ,  $D^2\phi$  and  $\phi_t$  are continuous in  $Q_T$ .

An equivalent definition of viscosity solutions will be utilized to establish the comparison principle for second order PDEs. To this end, it is necessary to prepare some notation: for  $u: \Omega \rightarrow \mathbf{R}$ , the semi-jets  $J^{2,\pm}u(x)$  of order 2 at  $x \in \Omega$  are defined by

$$J_{\Omega}^{2,+}u(x) = \left\{ (p, X) \in \mathbf{R}^n \times S^n \right. \\ \left. \begin{array}{l} u(y) \leq u(x) - \langle p, y - x \rangle \\ -\frac{1}{2} \langle X(y - x), y - x \rangle + o(|x - y|^2) \\ \text{for } y \in \Omega \text{ as } y \rightarrow x \end{array} \right\}, \\ J_{\Omega}^{2,-}u(x) = \left\{ (p, X) \in \mathbf{R}^n \times S^n \right. \\ \left. \begin{array}{l} u(y) \geq u(x) - \langle p, y - x \rangle \\ -\frac{1}{2} \langle X(y - x), y - x \rangle + o(|x - y|^2) \\ \text{for } y \in \Omega \text{ as } y \rightarrow x \end{array} \right\}.$$

In order to state a key lemma in Sect. “Comparison Principle”, it is also needed to introduce a sort of closures of  $J_{\Omega}^{2,\pm}u(x)$ ; for  $x \in \Omega$ ,

$$\bar{J}_{\Omega}^{2,\pm}u(x) = \left\{ (p, X) \in \mathbf{R}^n \times S^n \right. \\ \left. \begin{array}{l} \exists (p_k, X_k) \in J_{\Omega}^{2,\pm}u(x_k) \text{ for } \exists x_k \in \Omega, \\ \text{such that } (x_k, u(x_k), p_k, X_k) \\ \rightarrow (x, u(x), p, X) \text{ as } k \rightarrow \infty \end{array} \right\}.$$

**Definition 3 (Equivalent definition)**  $u$  is a viscosity subsolution (resp., supersolution) of (E) in  $\Omega$  if and only if

$$F(x, u(x), p, X) \leq 0 \quad (\text{resp., } \geq 0)$$

provided  $(p, X) \in \bar{J}_{\Omega}^{2,+}u(x)$  (resp.,  $(p, X) \in \bar{J}_{\Omega}^{2,-}u(x)$ ) for  $x \in \Omega$ .

One can replace  $\bar{J}_{\Omega}^{2,\pm}$  in the above by  $J_{\Omega}^{2,\pm}$ .

## Examples

Since the viscosity solution method is available only for second order elliptic/parabolic PDEs up to now, all the examples below are elliptic/parabolic.

To explain how the viscosity solution method is useful, a typical example is the eikonal equation: for  $\Omega_1 = (-1, 1) \subset \mathbf{R}$ ,

$$|Du| = 1 \quad \text{in } \Omega_1 \quad (2)$$

under the Dirichlet condition  $u(\pm 1) = 0$ . Since (2) is of non-divergence type, the notion of weak solutions in the distribution sense is not available to (2). It is also known that there are no classical solutions  $u \in C(\bar{\Omega}_1) \cap C^1(\Omega_1)$  of (2). In fact, from an optimal control and a geometric view points, the expected solution is the distance function from the boundary  $\partial\Omega_1$ ;  $u_0(x) = \text{dist}(x, \partial\Omega_1) = 1 - |x|$ .

If one employs Lipschitz continuous functions satisfying (2) almost every point of  $\Omega_1$  as weak solutions of (2), then the uniqueness of such weak solutions fails. Indeed, there exist many other weak solutions in this sense; e.g.

$$u_1(x) = |x| - 1, \\ u_2(x) = \begin{cases} \min\{x + 1, -x\} & (\text{for } x \in [-1, 1/2]), \\ x - 1 & (\text{for } x \in (1/2, 1]) \end{cases} \quad \text{etc.}$$

However, it can be proved that  $u_0$  is the unique viscosity solution of (2) under  $u_0(\pm 1) = 0$ . It is worth mentioning that there are no test functions  $\phi \in C^1(\Omega)$  touching  $u_0$  from below at  $x = 0$ . Thus, no inequalities for viscosity supersolutions at  $x = 0$  are required.

Moreover, if  $u_{\varepsilon}$  is a (classical) solution of the approximate equation via the vanishing viscosity method,

$$-\varepsilon \Delta u_{\varepsilon} + |Du_{\varepsilon}| = 1 \quad \text{in } \Omega_1 \quad (\varepsilon > 0),$$

under  $u_{\varepsilon}(\pm 1) = 0$ , then one can verify that  $\lim_{\varepsilon \rightarrow 0} u_{\varepsilon} = u_0$  in  $\bar{\Omega}_1$ .

The next example, to which the viscosity solution method is applicable, is a fully nonlinear PDE arising in stochastic control problems.

For a set of parameters  $\mathcal{A} \neq \emptyset$ ,  $F$  is expressed by

$$F(x, r, p, X) = \max_{a \in \mathcal{A}} \{ -\text{trace}(A_a(x)X) - \langle g_a(x), p \rangle + c_a(x)r - f_a(x) \},$$

where  $A_a: \Omega \rightarrow S^n$ ,  $g_a: \Omega \rightarrow \mathbf{R}^n$ ,  $c_a, f_a: \Omega \rightarrow \mathbf{R}$  are given functions for each  $a \in \mathcal{A}$ . One can see that the mapping  $(r, p, X) \rightarrow F(x, r, p, X)$  is convex in this case. The corresponding PDE is called the HJB equation:

$$\max_{a \in \mathcal{A}} \{ -\text{trace}(A_a(x)D^2u) - \langle g_a(x), Du \rangle + c_a(x)u - f_a(x) \} = 0. \quad (3)$$

If  $A_a \geq 0$  in  $\Omega$  for  $a \in \mathcal{A}$ , then (3) is elliptic.

From a stochastic control point of view,  $A_a(x)$  is naturally represented by  $1/2\sigma_a(x)\sigma_a^t(x)$  for an  $n \times m$  matrix  $\sigma_a(x)$  with some integer  $m$  (for  $a \in \mathcal{A}$  and  $x \in \Omega$ ). Thus, the property  $A_a \geq 0$  in  $\Omega$  holds true naturally in applications.

A fully nonlinear PDE for which the mapping  $(r, p, X) \rightarrow F(x, r, p, X)$  is neither convex nor concave arises in stochastic differential games:

$$\min_{b \in \mathcal{B}} \max_{a \in \mathcal{A}} \{ -\text{trace}(A_{a,b}(x)D^2u) - \langle g_{a,b}(x), Du \rangle + c_{a,b}(x)u - f_{a,b}(x) \} = 0,$$

where  $A_{a,b}: \Omega \rightarrow S^n$ ,  $g_{a,b}: \Omega \rightarrow \mathbf{R}^n$ ,  $c_{a,b}, f_{a,b}: \Omega \rightarrow \mathbf{R}$  are given for  $(a, b) \in \mathcal{A} \times \mathcal{B}$ . Here,  $\mathcal{B}$  is another set of parameters with a different purpose.

This is called the Hamilton–Jacobi–Bellman–Isaacs (abbreviated, HJBI) equation, for which the viscosity solution method works when  $A_{a,b} \geq 0$  in  $\Omega$  for  $(a, b) \in \mathcal{A} \times \mathcal{B}$ . In principle, any fully nonlinear second order PDEs can be represented by HJBI equations with suitable choices of  $A_{a,b}$ ,  $g_{a,b}$ ,  $c_{a,b}$ ,  $f_{a,b}$ ,  $\mathcal{A}$  and  $\mathcal{B}$  though those HJBI equations might become complicated.

It is trivial to derive parabolic versions of HJB and HJBI equations by adding  $u_t$  together with allowing  $t$ -dependence on the given functions.

There are more PDEs, to which the viscosity solution method has been applied.

$$\text{(Monge–Ampère equation)} \quad -\det(D^2u) - f(x) = 0,$$

$$\text{(mean curvature flow)} \quad u_t - \Delta u + \frac{\langle D^2u Du, Du \rangle}{|Du|^2} = 0,$$

$$\text{(\infty-Laplace equation)} \quad -\langle D^2u Du, Du \rangle = 0,$$

$$\text{(Pucci equations)} \quad \mathcal{P}^\pm(D^2u) - f(x) = 0.$$

The corresponding references are [26] for the Monge–Ampère equation, [23] for the mean curvature flow equation, [2,30] for  $\infty$ -Laplace equation, and [12,13] for Pucci equations.

### Comparison Principle

The comparison principle has been one of the main issues in the study of viscosity solutions since it implies their uniqueness.

The original idea to prove the comparison principle for first order PDEs is to double the number of variables with a clever choice of a penalization term. However, for second order PDEs, this technique did not directly work. The breakthrough was achieved by Jensen in [29].

The following lemma with matrix inequalities was originally formulated by Ishii.

**Lemma 4 (Ishii’s lemma)** *For  $u, v \in C(\Omega)$  and  $\phi \in C^2(\Omega \times \Omega)$ , if  $(x, y) \rightarrow u(x) - v(y) - \phi(x, y)$  takes its maximum at  $(\hat{x}, \hat{y}) \in \Omega \times \Omega$ , then for each  $\mu > 1$ , there are  $X, Y \in S^n$  such that  $(D_x\phi(\hat{x}, \hat{y}), X) \in \overline{J}_{\Omega}^{2,+}u(\hat{x})$ ,  $(-D_y\phi(\hat{x}, \hat{y}), Y) \in \overline{J}_{\Omega}^{2,-}v(\hat{y})$ , and*

$$-(\mu + \|A\|) \begin{pmatrix} I & O \\ O & I \end{pmatrix} \leq \begin{pmatrix} X & O \\ O & -Y \end{pmatrix} \leq A + \frac{1}{\mu}A^2, \quad (4)$$

where  $A = D^2\phi(\hat{x}, \hat{y}) \in S^{2n}$ .

The proof of this lemma consists of twice differentiability of convex functions by Aleksandrov, Aleksandrov–Bakelman–Pucci type maximum principle for semi-convex functions by Jensen, and the so-called magic properties of sup-convolutions.

For a typical  $\phi(x, y) = \frac{\mu}{2}|x - y|^2$ , the above matrix inequalities become

$$-3\mu \begin{pmatrix} I & O \\ O & I \end{pmatrix} \leq \begin{pmatrix} X & O \\ O & -Y \end{pmatrix} \leq 3\mu \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}. \quad (5)$$

In order to establish the comparison principle, the following two hypotheses on  $F$  are sufficient. The first one is monotonicity in the second variable:

there exists  $\theta > 0$  such that

$$r \rightarrow F(x, r, p, X) - \theta r \quad \text{is non-decreasing} \quad (6)$$

for  $(x, p, X) \in \Omega \times \mathbf{R}^n \times S^n$ . It should be noted that one may take  $\theta = 0$  in (6) if  $F$  is uniformly elliptic.

The next one is called the structure condition on  $F$ : by setting  $\mathcal{M} := \{\omega: [0, \infty) \rightarrow [0, \infty) | \omega \text{ is continuous such}$

that  $\omega(0) = 0$ ,

$$\begin{cases} \text{there exists } \omega \in \mathcal{M} \text{ such that } F(y, r, \mu(x-y), Y) \\ \leq F(x, r, \mu(x-y), X) + \omega(\mu|x-y|^2 + |x-y|) \\ \text{for } \mu > 1, x, y \in \Omega, r \in \mathbf{R}, X, Y \in S^n \text{ satisfying (5).} \end{cases} \quad (\text{SC})$$

Some observations related to (SC) below indicate that the condition (SC) is sufficiently reasonable.

#### Sufficient/Necessary Conditions for (SC):

- (i) If  $F$  satisfies (SC), then  $F$  is elliptic.
- (ii) If  $F$  is uniformly elliptic, and there is  $\bar{\omega} \in \mathcal{M}$  such that

$$|F(x, r, p, X) - F(y, r, p, X)| \leq \bar{\omega}(|x-y|(1+|p|+\|X\|))$$

for  $(x, y, r, p, X) \in \Omega \times \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n$ , then  $F$  satisfies (SC).

- (iii) In the case of HJB equations (3), if there are  $M_0 \geq 0$  and  $\omega_0 \in \mathcal{M}$  such that

$$\begin{cases} (a) \quad \max\{\|\sigma_a(x)\|, |g_a(x)|, |f_a(x)|, |c_a(x)|\} \leq M_0, \\ (b) \quad \max\{\|\sigma_a(x) - \sigma_a(y)\|, |g_a(x) - g_a(y)|\} \\ \leq M_0|x-y|, \\ (c) \quad \max\{|f_a(x) - f_a(y)|, |c_a(x) - c_a(y)|\} \\ \leq \omega_0(|x-y|) \end{cases} \quad (7)$$

for  $(x, y, a) \in \Omega \times \Omega \times \mathcal{A}$ , then  $F$  satisfies (SC).

It is obvious to give the sufficient condition corresponding to (7) for HJBI equations.

Due to the comparison principle below, it follows the uniqueness of viscosity solutions of (E) in  $\Omega$  such that  $u = h$  on  $\partial\Omega$ , where  $h \in C(\partial\Omega)$ .

**Comparison Principle:** Under hypotheses (6) and (SC), if  $u \in C(\bar{\Omega})$  and  $v \in C(\bar{\Omega})$  are, respectively, a viscosity subsolution and a viscosity supersolution of (E) in  $\Omega$ , then  $u \leq v$  on  $\partial\Omega$  yields  $u \leq v$  in  $\Omega$ .

In order to understand how Ishii's lemma is used for the proof of the comparison principle, a simple second order PDE with variable coefficients to the second derivatives will be chosen.

**Doubling the Number of Variables:** Lipschitz continuous functions  $\sigma_j = (\sigma_{ij}): \Omega \rightarrow \mathbf{R}^n$  ( $j = 1, \dots, m$ ) are given;  $|\sigma_j(x) - \sigma_j(y)| \leq M_0|x-y|$  ( $x, y \in \Omega$ ). By setting  $A(x) = \frac{1}{2} \sum_{k=1}^m \sigma_{ik}(x) \sigma_{jk}(x)$ , the following simple linear PDE is chosen as a typical example:

$$-\text{trace}(A(x)D^2u) + u = 0 \quad \text{in } \Omega.$$

Here  $\Omega$  is a bounded domain for simplicity. It is easy to verify that this PDE satisfies (6) and (SC).

A typical proof in viscosity solution theory is by contradiction. If  $\theta := \sup_{\bar{\Omega}}(u-v) > 0$ , and  $\hat{x} \in \bar{\Omega}$  is the point such that  $(u-v)(\hat{x}) = \theta$ , then  $\hat{x} \in \Omega$  because  $u \leq v$  on  $\partial\Omega$ . Here, a perturbation may allow  $\hat{x}$  to be the strict maximum. Now, the key idea is to consider the function  $(x, y) \in \bar{\Omega} \times \bar{\Omega} \rightarrow u(x) - v(y) - \frac{\mu}{2}|x-y|^2$  for  $\mu > 1$ . For  $(x_\mu, y_\mu) \in \bar{\Omega} \times \bar{\Omega}$  being a maximum of this function, one can first observe that

$$\lim_{\mu \rightarrow \infty} (x_\mu, y_\mu) = (\hat{x}, \hat{x}), \quad \text{and} \quad \lim_{\mu \rightarrow \infty} \mu|x_\mu - y_\mu|^2 = 0.$$

Due to Ishii's lemma, there exist  $X_\mu$  and  $Y_\mu \in S^n$  such that

$$\begin{pmatrix} X_\mu & O \\ O & -Y_\mu \end{pmatrix} \leq 3\mu \begin{pmatrix} I & -I \\ -I & I \end{pmatrix},$$

$(\mu(x_\mu - y_\mu), X_\mu) \in \bar{J}_{\Omega}^{2,+} u(x_\mu)$  and  $(\mu(x_\mu - y_\mu), Y_\mu) \in \bar{J}_{\Omega}^{2,-} v(y_\mu)$ . By setting  $\xi_j := (\sigma_{1j}(x_\mu), \dots, \sigma_{nj}(x_\mu))^t$  and  $\eta_j := (\sigma_{1j}(y_\mu), \dots, \sigma_{nj}(y_\mu))^t$  ( $j = 1, \dots, m$ ), the above matrix inequality implies

$$\begin{aligned} \left\langle \begin{pmatrix} X_\mu & O \\ O & -Y_\mu \end{pmatrix} \begin{pmatrix} \xi_j \\ \eta_j \end{pmatrix}, \begin{pmatrix} \xi_j \\ \eta_j \end{pmatrix} \right\rangle &\leq 3\mu|\xi_j - \eta_j|^2 \\ &\leq 3\mu M_0^2|x_\mu - y_\mu|^2. \end{aligned}$$

Thus, by summing these over  $j = 1, \dots, m$ , it follows that

$$\text{trace}(A(x_\mu)X_\mu) - \text{trace}(A(y_\mu)Y_\mu) \leq \frac{3}{2}\mu m M_0^2|x_\mu - y_\mu|^2.$$

Therefore, since the mappings  $x \rightarrow u(x) - v(y_\mu) - \frac{\mu}{2}|x - y_\mu|^2$  and  $y \rightarrow v(y) - u(x_\mu) + \frac{\mu}{2}|y - x_\mu|^2$ , respectively, take their maximum and minimum at  $x_\mu$  and  $y_\mu \in \Omega$  for large  $\mu > 1$ , the definition yields

$$\begin{aligned} u(x_\mu) - v(y_\mu) &\leq \text{trace}(A(x_\mu)X_\mu) - \text{trace}(A(y_\mu)Y_\mu) \\ &\leq \frac{3}{2}\mu m M_0^2|x_\mu - y_\mu|^2, \end{aligned}$$

which gives  $\theta \leq 0$  in the limit as  $\mu \rightarrow \infty$ . This is a contradiction.

#### Existence Results

There are two basic ways to show the existence of viscosity solutions. The first one is motivated by the classical Perron's method for harmonic functions.

For Perron's method below, it is necessary to introduce some relaxations of the notion. In the definition of viscosity subsolutions (resp., supersolution),  $u$  is replaced by  $u^*$



(resp.,  $u_*$ ), which is the upper (resp., lower) semicontinuous envelope of  $u$ ; for  $u: \Omega \rightarrow \mathbf{R}$  and  $x \in \Omega$ ,

$$u^*(x) = \lim_{\varepsilon \rightarrow 0} \sup_{\Omega \cap B_\varepsilon(x)} u \quad \left( \text{resp., } u_*(x) = \lim_{\varepsilon \rightarrow 0} \inf_{\Omega \cap B_\varepsilon(x)} u \right).$$

Here and later,  $B_\varepsilon(x)$  denotes the open ball with its center  $x \in \mathbf{R}^n$  and its radius  $\varepsilon > 0$ .

For the reader's convenience, the definition of viscosity sub- and supersolutions is restated.

**Definition 5**  $u: \Omega \rightarrow \mathbf{R}$  is called a viscosity subsolution (resp., supersolution) of (E) in  $\Omega$  if

$$\begin{aligned} F(x, u^*(x), D\phi(x), D^2\phi(x)) &\leq 0 \\ (\text{resp., } F(x, u_*(x), D\phi(x), D^2\phi(x)) &\geq 0) \end{aligned}$$

whenever  $(u^* - \phi)(x) = \sup_{\Omega} (u^* - \phi)$  (resp.,  $(u_* - \phi)(x) = \inf_{\Omega} (u_* - \phi)$ ) for  $\phi \in C^2(\Omega)$  and  $x \in \Omega$ .

It should be remarked that for the function  $u$ , no hypotheses are assumed in this definition.

**Remark 6** Ishii's lemma holds for  $u \in USC(\Omega)$  and  $v \in LSC(\Omega)$ , where for  $U \subset \mathbf{R}^n$ ,  $USC(U)$  and  $LSC(U)$  are, respectively, the sets of upper and lower semicontinuous functions in  $U$ . With these terminologies,  $u$  and  $v$  are, respectively, replaced by  $u^*$  and  $v_*$  in the comparison principle.

It is important to note that if the comparison principle holds for (E), and a viscosity solution  $u$  is continuous on  $\partial\Omega$ , then  $u$  is continuous in  $\overline{\Omega}$  even though  $u$  is not initially supposed to belong to  $C(\Omega)$ .

Ishii (cf. [27]) established Perron's method to show the existence of viscosity solutions.

**Perron's Method:** If there is a viscosity subsolution  $\xi \in USC(\overline{\Omega})$  and a viscosity supersolution  $\eta \in LSC(\overline{\Omega})$  of (E) in  $\Omega$  such that  $\xi \leq \eta$  in  $\overline{\Omega}$ , by setting

$$u(x) := \sup \left\{ v(x) \mid \begin{array}{l} v \in USC(\overline{\Omega}) \text{ is a viscosity subsolution} \\ \text{of (E) in } \Omega \text{ such that } \xi \leq v \leq \eta \text{ in } \overline{\Omega} \end{array} \right\},$$

then  $u$  is a viscosity solution of (E) in  $\Omega$ ;  $u^*$  and  $u_*$  are, respectively, a viscosity subsolution and a viscosity supersolution of (E) in  $\Omega$ .

The above  $u$  does not belong to  $USC(\Omega) \cup LSC(\Omega)$  in general.

The proof of Perron's method consists of two parts. One of them should be separately stated since this property is often useful.

**Proposition 7** For a non-empty set  $S \subset USC(\Omega)$  (resp.,  $LSC(\Omega)$ ) of viscosity subsolutions (resp., supersolutions) of (E) in  $\Omega$ , if  $u(x) := \sup\{v(x) \mid v \in S\}$  (resp.,  $u(x) := \inf\{v(x) \mid v \in S\}$ ) ( $x \in \Omega$ ) is locally bounded from above (resp., from below), then  $u$  is a viscosity subsolution (resp., supersolution) of (E) in  $\Omega$ .

**Remark 8** Thanks to Perron's method, for the existence of viscosity solutions, it is enough to construct a viscosity subsolution  $\xi$  and a viscosity supersolution  $\eta$  of (E) in  $\Omega$  satisfying  $\xi \leq \eta$  in  $\overline{\Omega}$ .

For instance, concerning the Dirichlet boundary value problem, for a given  $h \in C(\partial\Omega)$ , if there are  $\xi$  and  $\eta$  satisfying  $\xi \leq \eta$  in  $\Omega$ , and  $\xi = \eta = h$  on  $\partial\Omega$ , then the  $u$  constructed by Perron's method is a viscosity solution of (E) in  $\Omega$  under  $u = h$  on  $\partial\Omega$ . To find such a couple  $(\xi, \eta)$  of viscosity sub- and supersolutions, when  $F$  is uniformly elliptic and  $\partial\Omega$  satisfies the so-called uniform exterior cone condition, it is possible to build them by using a barrier function at each  $x_0 \in \partial\Omega$ .

For HJB/HJBI equations, there is the other approach for the existence result. Concerning HJB Eqs. (3) for instance, it is known that the expected solution is the value function associated with a stochastic control problem. To avoid the boundary condition, the domain  $\Omega$  for (3) is fixed by  $\mathbf{R}^n$ . For a given (progressively) measurable function  $\alpha: [0, \infty) \rightarrow \mathcal{A}$ , under assumption (7), it is known that one can solve the stochastic differential equation (abbreviated, SDE)

$$dX(t) = g_{\alpha(t)}(X(t))dt + \sigma_{\alpha(t)}(X(t))dW(t) \quad \text{for } t > 0,$$

under  $X(0) = x \in \mathbf{R}^n$ , where  $W$  is the  $m$ -dimensional Brownian motion in a probability space. By denoting the solution of the above by  $X(\cdot; x, \alpha)$ , a cost functional is given by

$$J(x, \alpha) = E \left[ \int_0^\infty e^{-\int_0^t c_{\alpha(s)}(X(s; x, \alpha))ds} \cdot f_{\alpha(t)}(X(t; x, \alpha))dt \mid X(0; x, \alpha) = x \right],$$

where  $E$  is the expectation. Then, the value function is defined by

$$u(x) = \inf_{\alpha} J(x, \alpha), \quad (8)$$

where the infimum is taken over all measurable controls.

The DPP is a formula which the value function satisfies when  $\inf\{c_a(x) \mid a \in \mathcal{A}, x \in \mathbf{R}^n\} > 0$ .

**Dynamic Programming Principle:** For  $\tau > 0$ ,

$$u(x) = \inf_{\alpha} E \left[ \int_0^{\tau} e^{-\int_0^t c_{\alpha(s)}(X(s; x, \alpha)) ds} f_{\alpha(t)}(X(t; x, \alpha)) dt + e^{-\int_0^{\tau} c_{\alpha(t)}(X(t; x, \alpha)) dt} u(X(\tau; x, \alpha)) \right] \Bigg| X(0; x, \alpha) = x.$$

Due to this DPP under (7), it is shown that  $u$  is a viscosity solution of (3) in  $\mathbf{R}^n$  via the Itô formula.

### Boundary Value Problems

In the study of degenerate elliptic/parabolic PDEs via viscosity solution method, there is a unified way to formulate boundary value problems of

$$\begin{cases} F(x, u, Du, D^2u) = 0 & \text{in } \Omega, \\ B(x, u, Du) = 0 & \text{on } \partial\Omega, \end{cases} \quad (9)$$

where  $B: \partial\Omega \times \mathbf{R} \times \mathbf{R}^n \rightarrow \mathbf{R}$  is a given (continuous) function.

Typical examples of  $B$  in the literature of PDEs are

$$\begin{cases} B(x, r, p) = r - h(x) & \text{(Dirichlet condition),} \\ B(x, r, p) = \langle p, \gamma(x) \rangle - h(x) & \text{(oblique condition),} \end{cases}$$

where  $h \in C(\partial\Omega)$ , and  $\gamma: \partial\Omega \rightarrow \mathbf{R}^n$  with  $\langle n(x), \gamma(x) \rangle > 0$  (for  $x \in \partial\Omega$ ). Here,  $n(x)$  is the outward unit normal at  $x \in \partial\Omega$ . When  $\gamma = n$ , the oblique condition is called the Neumann condition.

To formulate the boundary value problem (9) in the viscosity sense, by setting

$$G(x, r, p, X) = \begin{cases} F(x, r, p, X) & \text{(for } x \in \Omega), \\ B(x, r, p) & \text{(for } x \in \partial\Omega) \end{cases}$$

in the definition of viscosity subsolutions (resp., supersolutions),  $\Omega$  and  $F$  are replaced, respectively, by  $\overline{\Omega}$  and  $G_*$  (resp.,  $G^*$ ). Here,  $G^*$  and  $G_*$  are, respectively, an upper and lower semicontinuous envelopes of  $G$  with respect to all the variables.

When  $F$  is continuous,  $G^*(x, r, p, X) = G_*(x, r, p, X) = F(x, r, p, X)$  for  $(x, r, p, X) \in \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n$ . Moreover, if  $F$  is continuous up to  $\partial\Omega$  in the first variable, and  $B$  is continuous, then one can verify that at  $x \in \partial\Omega$ ,

$$\begin{aligned} G_*(x, r, p, X) &= \min\{F(x, r, p, X), B(x, r, p)\}, \\ G^*(x, r, p, X) &= \max\{F(x, r, p, X), B(x, r, p)\} \end{aligned}$$

for  $(r, p, X) \in \mathbf{R} \times \mathbf{R}^n \times S^n$ .

Under this setting, the comparison principle means that if  $u$  and  $v$  are, respectively, a viscosity subsolution and a viscosity supersolution of (9) in the above sense, then  $u \leq v$  in  $\overline{\Omega}$ . To prove the comparison principle for (9), the key tool is often a good perturbation of the penalization term to avoid the boundary condition.

If the boundary condition is imposed on the whole of  $\partial\Omega$ , then it would be an over-determined (i. e. unsolvable) problem when (E) is only elliptic. However, in the above setting, the given boundary condition,  $B(x, u, Du) = 0$  on  $\partial\Omega$ , is required only on a part of  $\partial\Omega$ , which is not a priori known.

This formulation has been successful particularly for oblique boundary value problems. For more information on the boundary value problems, [18] and its references are suggested.

**State Constraint Problems:** There is a sort of boundary condition arising in optimal control problems. However, it seems that this boundary condition is not derived from physical applications.

When the infimum in (8) is taken over all controls  $\alpha$  such that  $X(t; x, \alpha) \in \overline{\Omega}$  (when  $\overline{\Omega} \neq \mathbf{R}^n$ ) for all  $t \geq 0$ , it follows that the value function  $u$  is a viscosity supersolution of (3) in  $\overline{\Omega}$ .

Thus, to formulate the state constraint problem of (E), the supersolution property on  $\partial\Omega$  is regarded as a boundary condition;  $u: \overline{\Omega} \rightarrow \mathbf{R}$  is called a viscosity supersolution (resp., subsolution) of the state constraint problem of (E) in  $\Omega$  if it is a viscosity supersolution (resp., subsolution) of

$$F(x, u, Du, D^2u) = 0 \quad \text{in } \overline{\Omega} \quad (\text{resp., } \Omega).$$

The comparison principle for the state constraint problem was proved in [42] for first order PDEs.

### Asymptotic Analysis

Besides the existence and uniqueness of viscosity solutions, the remaining property in the well-posedness is the stability of those, which justifies that when the PDE is perturbed a little bit, the solution may change only a little.

In order to state the stability of viscosity solutions, Barles–Perthame in [8] introduced the half relaxed limits. For functions  $u_\varepsilon: \Omega \rightarrow \mathbf{R}$  for  $\varepsilon > 0$  and  $x \in \Omega$ ,  $\bar{u}(x) = \lim_{\varepsilon \rightarrow 0} \sup^* u_\varepsilon(x)$  (resp.,  $\underline{u}(x) = \lim_{\varepsilon \rightarrow 0} \inf_* u_\varepsilon(x)$ ) are defined by

$$\begin{aligned} &\lim_{\varepsilon \rightarrow 0} \sup \{u_\varepsilon(y) | 0 < \delta < \varepsilon, y \in B_\varepsilon(x)\} \\ &\left( \text{resp., } \lim_{\varepsilon \rightarrow 0} \inf \{u_\varepsilon(y) | 0 < \delta < \varepsilon, y \in B_\varepsilon(x)\} \right). \end{aligned}$$

For given  $F_\varepsilon: \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \rightarrow \mathbf{R}$  ( $\varepsilon > 0$ ), the stability of viscosity solutions of

$$F_\varepsilon(x, u, Du, D^2u) = 0 \quad \text{in } \Omega \quad (10)$$

is stated in the following manner.

**Stability:** If  $u_\varepsilon$  are viscosity subsolutions (resp., supersolutions) of (10) in  $\Omega$  for  $\varepsilon > 0$ , and  $\bar{u}$  (resp.,  $\underline{u}$ ) is finite in  $\Omega$ , then it is a viscosity subsolution (resp., supersolution) of

$$\underline{F}(x, u, Du, D^2u) = 0 \quad (\text{resp.}, \bar{F}(x, u, Du, D^2u) = 0) \quad \text{in } \Omega, \quad (11)$$

where  $\underline{F}(x, r, p, X)$  (resp.,  $\bar{F}(x, r, p, X)$ ) is given by

$$\liminf_{\varepsilon \rightarrow 0} \left\{ F_\delta(y, s, q, Y) \mid \begin{array}{l} y \in B_\varepsilon(x), |r-s| < \varepsilon, q \in B_\varepsilon(p), \\ \|X - Y\| < \varepsilon, 0 < \delta < \varepsilon \end{array} \right\} \\ \left( \text{resp.}, \limsup_{\varepsilon \rightarrow 0} \left\{ F_\delta(y, s, q, Y) \mid \begin{array}{l} y \in B_\varepsilon(x), |r-s| < \varepsilon, q \in B_\varepsilon(p), \\ \|X - Y\| < \varepsilon, 0 < \delta < \varepsilon \end{array} \right\} \right).$$

Thus, if  $\bar{F} = \underline{F}$ , and the (6) and (SC) holds for  $\bar{F}$ , then  $\bar{u} = \underline{u}$  on  $\partial\Omega$  yields  $\bar{u} = \underline{u}$  in  $\Omega$ , which implies the convergence of  $u_\varepsilon$  to  $\bar{u} = \underline{u}$ .

In order to understand how this stability is powerful, it is efficient to consider the singular perturbation of (E) with the vanishing viscosity term:

$$-\varepsilon \Delta u + F(x, u, Du, D^2u) = 0 \quad \text{in } \Omega, \quad (12)$$

where  $\varepsilon > 0$ , under some boundary condition. When  $F$  is not uniformly elliptic, it is often easier to solve (12) than (E).

If  $u_\varepsilon$  is a (classical) solution of (12) in  $\Omega$  for  $\varepsilon > 0$ , and the uniform bound of  $u_\varepsilon$  is known, then the above stability result together with the comparison principle for (E) implies the convergence of  $u_\varepsilon$ . On the other hand, if one follows the standard argument, then it is necessary to get higher a priori estimates on  $u_\varepsilon$  so that  $u_\varepsilon$  has a limit (along a subsequence if necessary).

In applications from science and engineering, there are various singular perturbation problems where  $\bar{F} \neq \underline{F}$ . One typical example is the following.

**Homogenization:** When the PDE is formulated by

$$F\left(\frac{x}{\varepsilon}, u_\varepsilon, Du_\varepsilon, D^2u_\varepsilon\right) = 0 \quad \text{in } \mathbf{R}^n,$$

where the mapping  $y \rightarrow F(y, r, p, X)$  satisfies the periodicity

$$F(y + z, r, p, X) = F(y, r, p, X)$$

for  $(y, z, r, p, X) \in \mathbf{R}^n \times \mathbf{Z}^n \times \mathbf{R} \times \mathbf{R}^n \times S^n$ , this PDE corresponds to the case when it has highly oscillating coefficients in a periodic medium.

Under certain hypothesis, e.g. coercivity for first order PDEs, uniform ellipticity for second order ones, or more general cases, one may show that a subsequence of  $u_\varepsilon$  converges to  $u$ , which is a viscosity solution of the effective PDE:

$$\hat{F}(u, Du, D^2u) = 0 \quad \text{in } \Omega,$$

which is called the cell problem. It is important to characterize  $\hat{F}$  in the homogenization theory since this new PDE may contain macroscopic information.

While there is extensive literature on homogenization for PDEs of divergence type, less was known for those of non-divergence type. The so-called perturbed test function method was proposed to study homogenization for non-divergent PDEs (see [16]). Now, even for PDEs of non-divergence type, the study of homogenization has developed in periodic/almost periodic environments.

Recently, the research of homogenization in ergodic media has started in [14,37].

There are also some singular perturbation problems in phase transitions. Section 5 in [5] is a good survey for those and their front propagations (see also [9]).

## Other Notions

There are some other notions of viscosity solutions. Indeed, although the viscosity solution method can be applied to various PDEs in a unified manner, one might lose some information in each specific problem.

**Semicontinuous Viscosity Solutions:** In the study of first order PDEs with convex Hamiltonian with respect to  $Du$ ,

$$H(x, u, Du) = 0 \quad \text{in } \Omega, \quad (13)$$

where  $H: \Omega \times \mathbf{R} \times \mathbf{R}^n \rightarrow \mathbf{R}$  is continuous, and  $p \rightarrow H(x, r, p)$  is convex for  $(x, r) \in \Omega \times \mathbf{R}$ , Barron and Jensen proposed the notion of semicontinuous viscosity solutions. In fact, it sometimes happens that the expected solution is discontinuous while the standard comparison principle implies its continuity. In such a problem, it is necessary to

propose a better notion of weak solutions under which one guarantees the uniqueness of discontinuous solutions.

**Definition 9**  $u \in USC(\Omega)$  is called a semicontinuous viscosity solution of (12) in  $\Omega$  if

$$H(x, \phi(x), D\phi(x)) = 0$$

whenever  $\sup_{\Omega}(u - \phi) = (u - \phi)(x)$  for  $\phi \in C^1(\Omega)$  and  $x \in \Omega$ .

It should be noted that the equality holds at  $x$  instead of the inequality in the standard definition.

It is known that under general setting, the value function in deterministic optimal control problems satisfies this definition. Moreover, the uniqueness of semicontinuous viscosity solutions is proved in [6].

**$L^p$ -viscosity Solutions:** In the study of (linear) uniformly elliptic/parabolic PDEs of divergence type, there are two basic regularity theories; the Schauder estimates and  $L^p$  estimates.

Initiated by Caffarelli in [11], the regularity theory of viscosity solutions for fully nonlinear uniformly elliptic/parabolic PDEs was developed. Afterwards, when  $F$  is uniformly elliptic but the mapping  $x \rightarrow F(x, r, p, X)$  is merely measurable, the notion of  $L^p$ -viscosity solutions of (E) in  $\Omega$  was introduced in [13], roughly speaking, by using  $W_{\text{loc}}^{2,p}(\Omega)$  instead of  $C^2(\Omega)$  in the definition for  $p > n/2$ .

**Definition 10**  $u \in C(\Omega)$  is an  $L^p$ -viscosity subsolution (resp., supersolution) of (E) in  $\Omega$  if

$$\begin{aligned} \text{ess lim inf}_{y \rightarrow x} F(y, u(y), D\phi(y), D^2\phi(y)) &\leq 0 \\ (\text{resp., } \text{ess lim sup}_{y \rightarrow x} F(y, u(y), D\phi(y), D^2\phi(y)) &\geq 0) \end{aligned}$$

whenever  $(u - \phi)(x) = \sup_{\Omega}(u - \phi)$  (resp.,  $\inf_{\Omega}(u - \phi)$ ) for  $\phi \in W_{\text{loc}}^{2,p}(\Omega)$  and  $x \in \Omega$ .

As usual,  $u \in C(\Omega)$  is called an  $L^p$ -viscosity solution of (E) in  $\Omega$  if it is an  $L^p$ -viscosity sub- and supersolution of (E) in  $\Omega$ .

A remarkable fact is that there are counter-examples even for linear uniformly elliptic PDEs for which uniqueness fails when the coefficient of the second derivatives is merely bounded. Thus, one may not expect the comparison principle to hold in general.

Here, several other notions are briefly mentioned.

Though the vanishing viscosity method works well for scalar conservation law, it is not clear if the viscosity solution method can be applied to it. There is a recent trial

in the study of first order PDEs of divergence/non-divergence type in [22]. The main idea for the new notion is to consider the associated level set equations, and to use test surfaces instead of test functions in the definition.

There are recent works on stochastic PDEs by introducing a stochastic version of viscosity solutions in [35,36].

## Future Directions

A list of areas where more research is expected via viscosity solution method is given.

**Venttsel condition:** There is a boundary condition where the second derivatives are involved;  $B$  contains  $D^2u$ -variables. In the case when  $X \rightarrow B(x, r, p, X)$  is elliptic, (9) with this  $B$  is called the Venttsel boundary value problem. From a view point of viscosity solution theory, this case corresponds to that of discontinuous coefficients.

**More regularity theory:** For fully nonlinear uniformly elliptic/parabolic PDEs, there are still many open questions on regularity.

The convexity (or concavity) assumption in the  $D^2u$ -variables on  $F$  seems necessary to show that viscosity solutions are classical ones. Indeed, some counter-examples have been discovered in [39,40] without convexity assumption. On the contrary, for a simple uniformly elliptic HJBI equation, it is shown that a viscosity solution is indeed a classical one in [10]. For degenerate elliptic but specified PDEs, there must be some criterion for regularity. For example, [41] for  $\infty$ -Laplace equation etc.

**Qualitative properties:** One may expect qualitative properties, which classical (or even strong) solutions possess, to hold for viscosity solutions such as Aleksandrov–Bakelman–Pucci maximum principle, Liouville theorem, Hadamard theorem, Phragmén–Lindelöf principle, symmetry etc. There have appeared several researches but there are still gaps to be fulfilled.

Concerning on the other qualitative properties, the most important one among those must be the eigenvalue problems but only a few results have appeared (e.g. [34]).

Also, the blow-up phenomenon for fully nonlinear parabolic PDEs is an attractive topic but almost nothing is known except [21].

**Degenerate equations from geometry:** There have been several results which treat degenerate second order PDEs arising in differential geometry; e.g. Heisenberg Laplacian in [4,38].

**Dynamical systems:** The weak KAM theory is an active area nowadays in viscosity solution theory. The part “Hamilton–Jacobi equations and weak KAM theory” in this section is recommended for more information.

**System of PDEs/nonlocal operators:** Viscosity solution method can naturally apply to monotone systems of PDEs in optimal control problems and differential games such as weakly coupled systems and switching games. Also, in applications, PDEs involving non-local operators such as impulsive control problems have been studied. There is a general treatment for PDEs with non-local operators in [28].

Recently, [1] investigated Hamilton–Jacobi equations with some special non-local operator which comes from a dislocation dynamics. One should seek new systems of PDEs, and PDEs with other non-local operators in various applications.

**Applications to mathematical finance:** There are lots of fully nonlinear PDEs appearing in mathematical finance. However, many of those cannot be applied directly by viscosity solution method since the PDEs contain unbounded coefficients, the domain is often unbounded, etc. Since the goal in this field is to determine (approximate) optimal policies, it is necessary to study the regularity of viscosity solutions in each problem.

There have also been several books in this field (e.g. [31]).

**Differential games:** Although HJBI equations appeared as examples from differential games, no details are mentioned. One interesting question in differential games is how to formulate the boundary condition for the state constraint problems. There was a trial [32] but it is not clear if this formulation covers practical problems.

**Control problems in  $\infty$ -dimensional spaces:** If (stochastic) PDEs govern the underlying state equation in place of (SDE) to define the value function in (8), then the domain of the resulting HJB equations becomes an  $\infty$ -dimensional space (typically, Hilbert space). The viscosity solution theory for  $\infty$ -dimensional spaces has been studied. However, much more research should be done for specific PDEs such as the Navier–Stokes equation (e.g. in [25]). Again [18] is suggested to find more information.

To finish this section, the reader should recall the words in the “User’s Guide” [18].

“the reader should not restrict his imagination to the borders we drew above.”

## Bibliography

### Primary Literature

1. Alvarez O, Hoch P, Le Bouar Y, Monneau R (2006) Dislocation dynamics: short time existence and uniqueness of the solution. *Arch Ration Mech Anal* 85:371–414
2. Aronsson G, Crandall MG, Juutinen P (2004) A tour of the theory of absolutely minimizing functions. *Bull Amer Math Soc* 41:439–505
3. Bardi M, Capuzzo Dolcetta I (1997) *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*. Birkhäuser, Boston
4. Bardi M, Mannucci P (2006) On the Dirichlet problem for non-totally degenerate fully nonlinear elliptic equations. *Comm Pure Appl Anal* 5:709–731
5. Bardi M, Crandall MG, Evans LC, Soner HM, Souganidis PE (1997) *Viscosity Solutions and Applications*. Springer, Berlin
6. Barles G (1993) Discontinuous viscosity solutions of first-order Hamilton–Jacobi equations: A guided visit. *Nonlinear Anal* 20:1123–1134
7. Barles G (1994) *Solutions de Viscosité des Équations de Hamilton–Jacobi*. Springer, Berlin
8. Barles G, Perthame B (1987) Discontinuous solutions of deterministic optimal stopping-time problems. *Model Math Anal Num* 21:557–579
9. Barles G, Souganidis PE (1998) A new approach to front propagation problems: theory and applications. *Arch Ration Mech Anal* 141:237–296
10. Cabré X, Caffarelli LA (2003) Interior  $C^{2,\alpha}$  regularity theory for a class of nonconvex fully nonlinear elliptic equations. *J Math Pures Appl* 83:573–612
11. Caffarelli LA (1989) Interior a priori estimates for solutions of fully non-linear equations. *Ann Math* 130:180–213
12. Caffarelli LP, Cabré X (1995) *Fully Nonlinear Elliptic Equations*. Amer Math Soc, Providence
13. Caffarelli LA, Crandall MG, Kocan M, Świąch A (1996) On viscosity solutions of fully nonlinear equations with measurable ingredients. *Comm Pure Appl Math* 49:365–397
14. Caffarelli LA, Souganidis PE, Wang L (2005) Homogenization of fully nonlinear, uniformly elliptic and parabolic partial differential equations in stationary ergodic media. *Comm Pure Appl Math* 58:319–361
15. Crandall MG, Lions P-L (1981) Condition d’unicité pour les solutions généralisées des équations de Hamilton–Jacobi du premier ordre. *CR Acad Sci Paris Sér I Math* 292:183–186
16. Crandall MG, Lions P-L (1983) Viscosity solutions of Hamilton–Jacobi equations. *Tran Amer Math Soc* 277:1–42
17. Crandall MG, Evans LC, Lions P-L (1984) Some properties of viscosity solutions of Hamilton–Jacobi equations. *Trans Amer Math Soc* 282:487–435
18. Crandall MG, Ishii H, Lions PL (1992) *User’s guide to viscosity solutions of second order partial differential equations*. *Bull Amer Math Soc* 27:1–67
19. Evans LC (1992) Periodic homogenization of certain fully nonlinear partial differential equations. *Proc Roy Soc Edinb* 120:245–265
20. Fleming WH, Soner HM (1993) *Controlled Markov Processes and Viscosity Solutions*. Springer, Berlin
21. Friedman A, Souganidis PE (1986) Blow-up of solutions of Hamilton–Jacobi equations. *Comm Partial Differ Equ* 11:397–443



22. Giga Y (2002) Viscosity solutions with shocks. *Comm Pure Appl Math* 55:431–480
23. Giga Y (2006) *Surface Evolutions Equations: A Level Set Approach*. Birkhäuser, Basel
24. Gilbarg D, Trudinger NS (1983) *Elliptic Partial Differential Equations of Second Order*. Springer, New York
25. Gozzi F, Sritharan SS, Święch A (2005) Bellman equations associated to optimal control of stochastic Navier–Stokes equations. *Comm Pure Appl Math* 58:671–700
26. Gutiérrez CE (2001) *The Monge–Ampère Equation*. Birkhäuser, Boston
27. Ishii H (1987) Perron’s method for Hamilton–Jacobi equations. *Duke Math J* 55:369–384
28. Ishii H, Koike S (1993/94) Viscosity solutions of functional-differential equations. *Adv Math Sci Appl* 3:191–218
29. Jensen R (1988) The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations. *Arch Rat Mech Anal* 101:1–27
30. Jensen R (1993) Uniqueness of Lipschitz extensions: minimizing the sup norm of the gradient. *Arch Ration Mech Anal* 123:51–74
31. Karatzas I, Shreve SE (1998) *Methods of Mathematical Finance*. Springer, New York
32. Koike S (1995) On the state constraint problem for differential games. *Indiana Univ Math J* 44:467–487
33. Koike S, Święch A (2004) Maximum principle for fully nonlinear equations via the iterated comparison function method. *Math Ann* 339:461–484
34. Lions P-L (1983) Bifurcation and optimal stochastic control. *Nonlinear Anal* 7:177–207
35. Lions PL, Souganidis PE (1998) Fully nonlinear stochastic partial differential equations. *CR Acad Sci Paris* 326:1085–1092; 326:753–741
36. Lions PL, Souganidis PE (2000) Uniqueness of weak solutions of fully nonlinear stochastic partial differential equations. *CR Acad Sci Paris* 331:783–790
37. Lions PL, Souganidis PE (2003) Correctors for the homogenization of Hamilton–Jacobi equations in the stationary ergodic setting. *Comm Pur Appl Math* 56:1501–1524
38. Manfredi JJ (2002) A version of the Hopf–Lax formula in the Heisenberg group. *Comm Partial Differ Equ* 27:1139–1159
39. Nadirashvili N, Vlăduț S (2007) Nonclassical solutions of fully nonlinear elliptic equations. *Geom Funct Anal* 17:1283–1296
40. Nadirashvili N, Vlăduț S (2008) Singular viscosity solutions to fully nonlinear elliptic equations. *J Math Pures Appl* 89(9):107–113
41. Savin O (2005)  $C^1$  regularity for infinity harmonic functions in two dimensions. *Arch Ration Mech Anal* 176:351–361
42. Soner MH (1986) Optimal control with state-space constraint I. *SIAM J Control Optim* 24:552–562
- Koike S (2004) *A Beginner’s Guide to the Theory of Viscosity Solutions*. Math Soc Japan, Tokyo. Corrections: <http://133.38.11.17/lab.jp/skoike/correction.pdf>
- Lions P-L (1982) *Generalized Solutions of Hamilton–Jacobi Equations*. Pitman, Boston
- Maugeri A, Palagachev DK, Softova LG (2000) *Elliptic and Parabolic Equations with Discontinuous Coefficients*. Wiley, Berlin

## Non-linear Stochastic Partial Differential Equations

GIUSEPPE DA PRATO

Scuola Normale Superiore, Pisa, Italy

### Article Outline

Glossary

Definition of the Subject

Introduction

General Problems and Results

Specific Equations

Future Directions

Bibliography

### Glossary

**Evolution equations** Let  $H$  be a Hilbert (or Banach) space,  $T > 0$  and  $f$  a mapping from  $[0, T] \times H$  into  $H$ . An equation of the form

$$u'(t) = f(t, u(t)) \quad , \quad t \in [0, T] \quad , \quad (*)$$

is called an *abstract evolution equation*. If  $H$  is finite-dimensional we call  $(*)$  an *ordinary evolution equation*, whereas if  $H$  is infinite-dimensional and  $f$  is a differential operator we call  $(*)$  a *partial differential evolution equation*.

**Continuous stochastic process** Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space. A *continuous stochastic process* (with values in  $H$ ) is a family of ( $H$ -valued) random variables  $(X(t) = X(t, \omega))_{t \geq 0}$  ( $\omega \in \Omega$ ) such that  $X(\cdot, \omega)$  is continuous for  $\mathbb{P}$ -almost all  $\omega \in \Omega$ .

**Brownian motion** A real *Brownian motion*  $B = (B(t))_{t \geq 0}$  on  $(\Omega, \mathcal{F}, \mathbb{P})$  is a continuous real stochastic process in  $[0, +\infty)$  such that (1)  $B(0) = 0$  and for any  $0 \leq s < t$ ,  $B(t) - B(s)$  is a real Gaussian random variable with mean 0 and covariance  $t - s$ , and (2) if  $0 < t_1 < \dots < t_n$ , the random variables,  $B(t_1)$ ,  $B(t_2) - B(t_1)$ ,  $\dots$ ,  $B(t_n) - B(t_{n-1})$  are independent.

### Books and Reviews

- Bellman R (1957) *Dynamic Programming*. Princeton University Press, Princeton
- Evans LC, Gangbo W (1999) *Differential Equations Methods for the Monge–Kantorovich Mass Transfer Problem*. Amer Math Soc, Providence
- Fleming WH, Rishel R (1975) *Deterministic and Stochastic Optimal Control*. Springer, New York

**Cylindrical Wiener process** A cylindrical Wiener process in a Hilbert space  $H$  is a process of the form

$$W(t) = \sum_{k=1}^{\infty} e_k \beta_k(t), \quad t \geq 0,$$

where  $(e_k)$  is a complete orthonormal system in  $H$  and  $(\beta_k)$  a sequence of mutually independent standard Brownian motions in a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ .

**Noise** Let  $W(t)$  be a cylindrical Wiener process in a Hilbert space  $H$ . A noise is an expression of the form  $BW(t)$ , where  $B \in L(H)$ . If  $B = I$  (the identity operator in  $H$ ) the noise is called *white*.

**Stochastic dynamical system** This is a system governed by a partial differential evolution equation perturbed by noise.

### Definition of the Subject

Several nonlinear partial differential evolution equations can be written as abstract equations in a suitable Hilbert space  $H$  (we shall denote by  $|\cdot|$  the norm and by  $\langle \cdot, \cdot \rangle$  the scalar product in  $H$ ) of the following form [78]:

$$\begin{cases} \frac{dX(t)}{dt} = AX(t) + b(X(t)), & t \geq 0, \\ X(0) = x \in H, \end{cases} \quad (1)$$

where  $A: D(A) \subset H \rightarrow H$  is the infinitesimal generator of a strongly continuous semigroup  $e^{tA}$  of linear bounded operators in  $H$  and  $b: D(b) \subset H \rightarrow H$  is a nonlinear mapping.

In order to take into account unpredictable random perturbations, one is led to add to (1) a term of the form  $\sigma(X(t))dW(t)$ , where  $\sigma: D(\sigma) \subset H \rightarrow L(H)$  ( $L(H)$  is the space of all linear operators from  $H$  into itself) and  $W(t)$  a cylindrical Wiener process in  $H$ . Then (1) is replaced by the following stochastic partial differential equation (SPDEs)

$$\begin{cases} dX(t) = (AX(t) + b(X(t)))dt + \sigma(X(t))dW(t), \\ X(0) = x \in H, \end{cases} \quad (2)$$

whose precise meaning will be explained later.

The importance of the SPDEs is due to the fact that in the modelization of several phenomena, (2) is often more realistic than (1).

The study of SPDEs started in the 1970s and 1980s and interest is still growing. Among the early contributions we mention [7] for Navier–Stokes equations, [66] for SPDEs

of monotone type, [79] for compactness methods, [57] for variational methods and [31].

Two approaches (essentially equivalent) are used for studying SPDEs: the first is based on the variational theory of partial differential equations and the second is based on the theory of strongly continuous semigroups of operators. In this article we are concerned with the latter [26,27]. For a recent introduction to the variational method see [71].

It is possible to consider perturbations of a partial differential equation by stochastic processes different from Brownian motion as jump processes, but we shall not consider this subject here.

### Introduction

In this section we are concerned with (2) in the Hilbert space  $H$ , that is, with a partial differential equation of the form (1) perturbed by the noise  $\sigma(X(t))dW(t)$ , where  $W(t)$  is a cylindrical Wiener process in  $H$ . We recall that  $W(t)$  can be defined as follows:

$$W(t) = \sum_{k=1}^{\infty} e_k \beta_k(t), \quad (3)$$

where  $e_k$  is a complete orthonormal system in  $H$  (often the system of eigenfunctions of  $A$ ) and  $\beta_k$  is a sequence of mutually independent standard Brownian motions in a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Intuitively  $W(t)$  represents a random (white) perturbation which has the same intensity in all direction of the orthonormal basis  $(e_k)$ .

**Remark 1** Definition (3) is formal since

$$\mathbb{E}[|W(t)|^2] = \sum_{k=1}^{\infty} t = +\infty \quad \forall t > 0.$$

However, it is possible to define rigorously  $W(t)$  in a suitable space larger than  $H$  [26]. We shall see in the following how to deal with  $W(t)$ .

The main problems we shall consider are the following:

- (i) Existence and uniqueness of solutions  $X(t)$  of (2).
- (ii) Asymptotic behavior of  $X(t)$  as  $t \rightarrow \infty$ .

For the sake of brevity we shall limit ourselves to stochastic perturbations of the special form  $\sigma(X(t))dW(t) = BdW(t)$ , where  $B \in L(H)$ . In this case we say that the noise involved in (2) is *additive*. The general case (*multiplicative noise*) is important and we shall give some references later. In particular a multiplicative noise is required when for physical reasons the solution  $X(t)$  must belong to some convex subset of  $H$  (for instance,  $X \geq 0$ ) [6].

In Sect. “General Problems and Results” we shall present an overview of the main problems concerning SPDEs. Section “Specific Equations” is devoted to some significant applications. We will not give complete proofs of all assertions but only a sketch in order to give an idea of the tools which are involved. The interested reader can look at the corresponding references.

We shall concentrate on some SPDEs only. We present in the following three examples the corresponding deterministic problems.

*Example 2* (reaction–diffusion equations). Consider the following partial differential equation in a bounded subset  $\mathcal{O}$  of  $\mathbb{R}^N$  with regular boundary  $\partial\mathcal{O}$ :

$$\begin{cases} \partial_t X(t, \xi) = \Delta_\xi X(t, \xi) + p(X(t, \xi)), & \xi \in \mathcal{O}, t > 0, \\ X(t, \xi) = 0, & t > 0, \xi \in \partial\mathcal{O}, \\ X(0, \xi) = x(\xi), & \xi \in \mathcal{O}, x \in H, \end{cases} \quad (4)$$

where  $\Delta_\xi$  is the Laplace operator and  $p$  is a polynomial of degree  $d > 1$  with negative leading coefficient.

Let us write problem (4) as an abstract equation of the form (1) in the Hilbert space  $H = L^2(\mathcal{O})$  (the space of all equivalence classes of square integrable real functions on  $\mathcal{O}$  with respect to the Lebesgue measure).

For this purpose, we denote by  $A$  the realization of the Laplace operator with Dirichlet boundary conditions (other boundary conditions such as Neumann or Ventzell could be considered as well),

$$\begin{cases} Ax = \Delta_\xi x, & x \in D(A), \\ D(A) = H^2(\mathcal{O}) \cap H_0^1(\mathcal{O}), \end{cases}$$

and define

$$b(x)(\xi) = p(x(\xi)), \quad x \in D(b) = L^{2d}(\mathcal{O}).$$

For  $k = 1, 2$  we denote by  $H^k(\mathcal{O})$  the Sobolev space consisting of all functions which belong to  $L^2(\mathcal{O})$  together with their distributional derivatives of order up to  $k$ .  $H_0^1(\mathcal{O})$  is the subspace of those functions in  $H^1(\mathcal{O})$  which vanish at the boundary  $\partial\mathcal{O}$  of  $\mathcal{O}$ .

*Example 3* (Burgers equation) Consider the equation in  $[0, 2\pi]$  with periodic boundary conditions

$$\begin{cases} \partial_t X(t, \xi) = \partial_\xi^2 X(t, \xi) + \frac{1}{2} \partial_\xi (X^2(t, \xi)), & \xi \in [0, 2\pi], t > 0, \\ X(t, 0) = X(t, 2\pi), \quad \partial_\xi X(t, 0) = \partial_\xi X(t, 2\pi), & t > 0, \\ X(0, \xi) = x(\xi), & \xi \in [0, 2\pi]. \end{cases} \quad (5)$$

Problem (5) can be written in the abstract form (1) setting  $H = L^2(0, 2\pi)$ ,

$$Ax = D_\xi^2 x, \quad x \in D(A) = H_\#^2(0, 2\pi),$$

and

$$b(x) = \frac{1}{2} D_\xi(x^2), \quad x \in D(b) = H_\#^1(0, 2\pi),$$

where

$$H_\#^1(0, 2\pi) = \{x \in H^1(0, 2\pi) : x(0) = x(2\pi)\}$$

and

$$\begin{aligned} H_\#^2(0, 2\pi) = \\ \{x \in H^2(0, 2\pi) : x(0) = x(2\pi), \partial_\xi x(0) = \partial_\xi x(2\pi)\}. \end{aligned}$$

*Example 4* (2D Navier–Stokes equation). Consider the equation in the square  $\mathcal{O} := [0, 2\pi] \times [0, 2\pi]$ .

$$\begin{cases} dZ = (\Delta_\xi Z - Z + D_\xi Z \cdot Z)dt + \nabla p dt & \xi \in \mathcal{O}, t > 0, \\ \operatorname{div} Z = 0 & \xi \in \mathcal{O}, t > 0, \\ Z(t, \cdot) \text{ is periodic with period } 2\pi, \\ Z(0, \xi) = z(\xi) & \xi \in \mathcal{O}. \end{cases} \quad (6)$$

The unknown  $Z = (Z_1, Z_2)$  represents the velocity and  $p$  the pressure of the fluid;  $\operatorname{div} Z$  is the divergence of  $Z$ . We denote by  $H$  the space of all square integrable divergence free vectors,

$$H = \{Z = (Z_1, Z_2) \in (L^2(0, 2\pi))^2 : \operatorname{div} Z = 0\},$$

and by  $\mathcal{P}$  the orthogonal projector of  $(L^2(0, 2\pi))^2$  onto  $H$ .

Then setting  $X(t, x) = \mathcal{P}Z(t, x)$ , we obtain the following problem (where the pressure  $p$  has disappeared)

$$\begin{cases} dX = \mathcal{P}(\Delta X - X) + \mathcal{P}(D_\xi X \cdot X) & \xi \in \mathcal{O}, t > 0, \\ X(t, \xi) \text{ is periodic with period } 2\pi, & t > 0 \\ X(0, \xi) = x & \xi \in \mathcal{O}. \end{cases}$$

Let us define the Stokes operator  $A: D(A) \rightarrow H$  setting

$$Ax = \mathcal{P}(\Delta_\xi x - x), \quad x \in D(A) = H_\#^2(\mathcal{O})$$

and the nonlinear operator  $b$  setting

$$b(x) = \mathcal{P}(D_\xi x \cdot x), \quad x \in H_\#^1(\mathcal{O})$$

( $b$  is more commonly written  $b(x) = \mathcal{P}(x \cdot \nabla x)$ ). Here  $H_\#^k(\mathcal{O})$ ,  $k = 1, 2$  are Sobolev spaces with periodic boundary conditions. Now, problem (5) can be written in the form (1).

We have to say that many other interesting equations have been studied recently. We mention among them the following (without claiming to present an exhaustive list).

- Cahn–Hilliard equations [21,38].
- Second-order linear SPDEs and Filtering equations; see [56,74] and references therein.
- Ginzburg–Landau equations [51,52,53].
- Kortweg–de Vries equation [36,37].
- Mathematical biology and Fleming–Viot model [48, 77].
- Mathematical finance [64].
- Nonlinear Schrödinger equations [32,33,34,35].
- Porous media equations [4,30,73].
- Stochastic quantization [2,23,25,55,59,62].
- Wave equations [5,18,19,62,63,69,75].

## General Problems and Results

### Well-Posedness of a Stochastic Partial Differential Equation

We are here concerned with (2). The first problem is to prove the existence and uniqueness of solutions (to be defined in an appropriate way). We notice that it is not convenient to look for existence of strong solutions  $X(t)$  (as for the ordinary stochastic equations) such that

$$X(t) = x + \int_0^t [AX(s) + b(X(s))] ds + BW(t), \quad t \geq 0.$$

This definition will require too much regularity for the solution, which does not belong to  $D(A) \cap D(b)$  in general. The concept of a weaker solution has to be defined for each specific problem; see Sect. “Specific Equations”.

Assume now that the existence of a unique solution  $X(t, x)$  (in a suitable sense) of (2) has been proved for any  $x \in H$ . Then, given a real function  $\varphi$  in  $H$ , one is interested in the evolution in time of  $\mathbb{E}[\varphi(X(t, x))]$  (the expectation of  $\varphi(X(t, x))$ ).  $\varphi$  can be interpreted as an *observable*; its physical meaning could be, for instance, the mean velocity, pressure or temperature of a fluid. This leads to introduce the concept of *transition semigroup*

$$P_t \varphi(x) = \mathbb{E}[\varphi(X(t, x))], \quad t \geq 0, \quad x \in H, \quad \varphi \in B_b(H), \quad (7)$$

where  $B_b(H)$  is the space of all mappings  $\varphi: H \rightarrow \mathbb{R}$  which are bounded and Borel. It is a Banach space with the supremum norm

$$\|\varphi\|_\infty = \sup_{h \in H} |\varphi(h)|, \quad \varphi \in B_b(H).$$

We denote by  $C_b(H)$  the closed subspace of  $B_b(H)$  of all uniformly continuous and bounded functions.

It is not difficult to see that  $P_t$  enjoys the semigroup property  $P_{t+s} = P_t P_s$ ,  $t, s \geq 0$ , thanks to the Markovianity of the solution  $X(t, x)$  to (2). However,  $P_t$  is not a strongly continuous semigroup on  $C_b(H)$  in general (even when  $H$  is finite-dimensional and (2) is an ordinary differential equation).

Formally, the function  $u(t, x) = P_t \varphi(x)$  is the unique solution to the following (Kolmogorov) parabolic equation:

$$\begin{cases} \frac{du}{dt}(t, x) = K_0 u(t, x), \\ u(0, x) = \varphi(x), \end{cases} \quad (8)$$

where  $K_0$  is the linear differential operator

$$K_0 \varphi(x) = \frac{1}{2} \text{Tr} [BB^* D_x^2 \varphi(x)] + \langle Ax + b(x), D_x \varphi(x) \rangle, \quad x \in D(A) \cap D(b), \quad (9)$$

where

$$\text{Tr} [BB^* D_x^2 \varphi(x)] = \sum_{h,k=1}^{\infty} \langle D_x^2 \varphi(x) B^* e_k, B^* e_k \rangle,$$

and  $(e_k)$  is a complete orthonormal system on  $H$ . Notice that  $K_0 \varphi$  is only defined for  $\varphi$  of class  $C^2$  such that the series above is convergent.

Though the semigroup  $P_t$  is not strongly continuous one can define its infinitesimal generator  $K$ , following [72],

$$K\varphi(x) = \lim_{h \rightarrow 0} \frac{1}{h} (P_t \varphi(x) - \varphi(x)), \quad x \in H,$$

provided the limit exists and the following condition holds

$$\sup_{h \in (0,1]} \frac{1}{h} \|P_t \varphi - \varphi\|_\infty < +\infty.$$

Then an interesting problem consists in clarifying the relationship between the abstract operator  $K$  and the concrete differential operator  $K_0$ .

Let us list some important concepts related to  $P_t$ .

- $P_t$  is called *Feller* if  $P_t \varphi \in C_b(H)$  for any  $\varphi \in C_b(H)$  and any  $t \geq 0$ .
- $P_t$  is called *strong Feller* if  $P_t \varphi \in C_b(H)$  for any  $\varphi \in B_b(H)$  and any  $t > 0$ .

The strong Feller property is related to a smoothing effect of the transition semigroup  $P_t$  and to the hypoellipticity of the Kolmogorov operator  $K_0$  (when  $H$  is finite-dimensional) [75].

- $P_t$  is called *irreducible* if  $P_t 1_I(x) > 0$  for all  $x \in H$  and all open sets  $I$  of  $H$ , where  $1_I$  is the characteristic function of  $I$  ( $1_I(x) = 1$  if  $x \in I$ ,  $1_I(x) = 0$  if  $x \notin I$ ).

Irreducibility of  $P_t$  is related to the null-controllability of the deterministic system (that is, if for each  $x \in H$  and  $T > 0$  there is a control  $u$  such that  $x(T) = 0$ ).

$$\begin{cases} \frac{dx(t)}{dt} = Ax(t) + b(x(t)) + Bu(t), & t \geq 0, \\ x(0) = x \in H, \end{cases}$$

where  $u$  is a control; see, e. g. Sects. 7.3 and 7.4 in [27].

Finally, we denote by  $\pi_t(x, dy)$  the law of  $X(t, x)$ , so that the following integral representation for  $P_t$ ,

$$P_t \varphi(x) = \int_H \varphi(y) \pi_t(x, dy), \quad \forall t > 0, x \in H, \varphi \in B_b(H), \quad (10)$$

holds.

### Asymptotic Behavior of the Transition Semigroup

In order to study the asymptotic behavior of  $P_t \varphi$ , where  $\varphi \in B_b(H)$ , the concept of *invariant measure* is crucial. We say that a Borel probability measure  $\nu$  on  $H$  is invariant for  $P_t$  if

$$\int_H P_t \varphi(x) \nu(dx) = \int_H \varphi(x) \nu(dx), \quad \forall \varphi \in C_b(H).$$

Notice that if  $\nu$  is invariant and  $\eta$  is a random variable whose law is  $\nu$  then the solution  $X(t, \eta)$  to (2) with initial datum  $\eta$  is stationary.

In order to prove the existence of invariant measures an important tool is the following Krylov–Bogoliubov theorem; see, e. g., Theorem 3.1.1 in [27].

**Theorem 5** Assume that  $P_t$  is Feller and that for some  $x \in H$  the family of probability measures  $(\pi_t(x, dy))_{t \geq 0}$  is tight (that is, for any  $\epsilon > 0$  there exists a compact set  $K_\epsilon$  in  $H$  such that  $\pi_t(x, K_\epsilon) > 1 - \epsilon$  for all  $t \geq 0$ ). Then there exists an invariant measure for  $P_t$ .

Assume now that  $\nu$  is an invariant measure for  $P_t$  and let  $\varphi \in C_b(H)$ . Then  $P_t$  can be uniquely extended to a contraction semigroup in  $L^2(H, \nu)$ . In fact, by (10) and the Hölder inequality we have

$$[P_t \varphi(x)]^2 \leq P_t(\varphi^2)(x), \quad t \geq 0, x \in H.$$

Integrating this inequality with respect to  $\nu$  over  $H$  and taking into account that  $\int_H P_t(\varphi^2)(x) \nu(dx) = \int_H \varphi^2(x) \nu(dx)$  by the invariance of  $\nu$  yields

$$\int_H [P_t \varphi(x)]^2 \nu(dx) \leq \int_H \varphi^2(x) \nu(dx).$$

Since  $C_b(H)$  is dense in  $L^2(H, \nu)$ , this inequality can be extended to any function of  $L^2(H, \nu)$ . This proves that  $P_t$  can be uniquely extended to a contraction semigroup in  $L^2(H, \nu)$ .

Moreover, the following Von Neumann theorem holds [70].

**Theorem 6** Let  $\nu$  be an invariant measure for  $P_t$ . Then for any  $\varphi \in L^2(H, \nu)$  there exists the limit

$$\lim_{T \rightarrow +\infty} \frac{1}{T} \int_0^T P_t \varphi(x) dt := \Pi \varphi,$$

where  $\Pi$  is a projection operator on

$$\Sigma := \{\varphi \in L^2(H, \nu) : P_t \varphi = \varphi, \forall t \geq 0\}.$$

If moreover

$$\lim_{T \rightarrow +\infty} \frac{1}{T} \int_0^T P_t \varphi(x) dt = \int_H \varphi(y) \nu(dy) \quad \text{in } L^2(H, \nu),$$

then  $\nu$  is called *ergodic*.

As in the deterministic case, ergodicity is a very important property of a stochastic dynamical system because it allows us to compute averages with respect to  $t$  in terms of averages with respect to  $x \in H$  (“temporal” averages with “spatial” averages).

When the more stringent condition

$$\lim_{T \rightarrow +\infty} P_t \varphi(x) = \int_H \varphi(y) \nu(dy), \quad \forall x \in H,$$

holds, we say that the measure  $\nu$  is *strongly mixing*. In this direction we recall the following Doob’s theorem; see e. g., Theorem 4.2.1 in [27].

**Theorem 7** Assume that  $P_t$  is irreducible and strongly Feller. Then  $P_t$  possesses at most one invariant measure which is ergodic and strongly mixing.

### Other Important Problems

Here we mention some additional problem which we cannot treat for space reasons. The first one concerns the so-called *small noise*. Often one is interested in studying stochastic perturbations of (1) of the form

$$\begin{cases} dX(t) = (AX(t) + b(X(t)))dt + \epsilon dW(t), \\ X(0) = x \in H, \end{cases} \quad (11)$$



where  $\epsilon > 0$ . In fact it can happen that (11) has a unique invariant measure  $\mu_\epsilon$ , whereas the corresponding deterministic system (1) has more. Then studying the limit points of  $\mu_\epsilon$  as  $\epsilon \rightarrow 0$  will help to select the “significant” invariant measures of (2) [39]. This problem also leads to study *large deviations* problems. Concerning Smoluchowski–Kramers approximations for parabolic SPDEs we mention [17].

## Specific Equations

### Ornstein–Uhlenbeck Equations

We are here concerned with the following stochastic differential equation:

$$\begin{cases} dX(t) = AX(t)dt + BdW(t), & t \geq 0 \\ X(0) = x \in H, \end{cases} \quad (12)$$

where  $A: D(A) \subset H \rightarrow H$  is the infinitesimal generator of a strongly continuous semigroup  $e^{tA}$  and  $B: H \rightarrow H$  is linear-bounded.

Formally the solution of (12) is given by the variation of constants formula

$$X(t, x) = e^{tA}x + W_A(t), \quad t \geq 0, \quad (13)$$

where the process  $W_A(t)$ , called *stochastic convolution*, is given by

$$W_A(t) = \int_0^t e^{(t-s)A} B dW(s), \quad t \geq 0.$$

By a formal computation we see that

$$\mathbb{E}|W_A(t)|^2 = \sum_{k=1}^{\infty} \int_0^t |e^{(t-s)A} B e_k|^2 ds = \text{Tr } Q_t,$$

where

$$Q_t x = \int_0^t e^{sA} C e^{sA*} x ds, \quad x \in H, \quad t \geq 0,$$

and  $C = BB^*$ . Consequently, though the definition (3) of the cylindrical Wiener process  $W(t)$  is formal (see Remark 1), the stochastic convolution defined by the series

$$W_A(t) = \sum_{k=1}^{\infty} \int_0^t e^{(t-s)A} B e_k d\beta_k(s) \quad (14)$$

is meaningful and convergent in  $L^2(\Omega, \mathcal{F}, \mathbb{P}; H)$  for all  $t \geq 0$ , provided  $\text{Tr } Q_t < +\infty$  for all  $t > 0$ . We shall assume that this hypothesis holds from now on.

It is easy to see that  $W_A(t)$  is a Gaussian random variable  $N_{Q_t}$  in  $H$ . (A Borel probability measure  $\mu$  in  $H$  is said

to be *Gaussian* with *mean*  $x \in H$  and *covariance*  $Q$ , where  $Q \in L(H)$  is of trace class, if the Fourier transform  $\hat{\mu}$  of  $\mu$  is given by  $\hat{\mu}(h) = e^{i\langle x, h \rangle - \frac{1}{2} \langle Qh, h \rangle}$  for all  $h \in H$ . In this case we set  $\mu = N_{x, Q}$  and  $\mu = N_Q$  if  $x = 0$ .)

**Remark 8** The limit case when  $B = I$  (*white noise*) [80], is important in some application in physics, because it means that the noise acts uniformly in all directions. In this case the assumption  $\text{Tr } Q_t < +\infty$  is equivalent to

$$\int_0^t \text{Tr} [e^{sA} e^{sA*}] ds < +\infty. \quad (15)$$

Assume, for instance, that  $A$  is the Laplacian in an open bounded set  $\mathcal{O} \subset \mathbb{R}^d$  with Dirichlet boundary conditions. Let  $Ae_k = \alpha_k e_k$ , where  $(e_k)_{k \in \mathbb{N}^d}$  is the complete orthonormal system of eigenfunctions of  $A$  and  $(\alpha_k)_{k \in \mathbb{N}^d}$  are the corresponding eigenvalues. Then we have

$$\text{Tr } Q_t = \sum_{k \in \mathbb{N}^d} \frac{1}{\alpha_k} (1 - e^{-2\alpha_k t}), \quad t > 0.$$

It is well known that  $\alpha_k \sim |k|^2$  and so (15) is fulfilled only if  $d = 1$ . To deal with the case  $d > 1$  one introduces the so-called *renormalization* [68, 76].

Now by (13) it follows that the law of  $X(t, x)$  is given by  $N_{e^{tA}x, Q_t}$ , so the corresponding transition semigroup looks like

$$P_t \varphi(x) = \int_H \varphi(y) N_{e^{tA}x, Q_t}(dy), \quad \varphi \in B_b(H).$$

If  $e^{tA}$  is stable,

$$\|e^{tA}\| \leq M e^{-\omega t}, \quad t \geq 0,$$

and for some  $M, \omega > 0$  it follows that

$$\lim_{t \rightarrow +\infty} P_t \varphi(x) = \int_H \varphi(y) N_{Q_\infty}(dy),$$

and so  $\mu := N_{Q_\infty}$  is the unique invariant measure for  $P_t$  which is in addition ergodic and strongly mixing (for a necessary and sufficient condition for existence and uniqueness of an invariant measure for  $P_t$  see [26]).

### Smooth Perturbations of System (12)

Let us consider the following stochastic differential equation

$$\begin{cases} dX(t) = (AX(t) + b(X(t)))dt + BdW(t), & t \geq 0, \\ X(0) = x \in H, \end{cases} \quad (16)$$

where  $A$  and  $B$  are as before and  $b: H \rightarrow H$  is of class  $C^1$  and Lipschitz continuous. The condition on  $b$  is very strong; however, this case is important because it can be used for approximating equations with irregular coefficients.

By a *mild* solution of problem (16) on  $[0, T]$  we mean a stochastic process  $X \in C_W([0, T]; H)$  such that

$$X(t) = e^{tA}x + \int_0^t e^{(t-s)A}b(X(s))ds + W_A(t), \quad \mathbb{P}\text{-a.s.} \quad (17)$$

where  $W_A$  is the stochastic convolution defined by (14). By  $C_W([0, T]; H)$  we mean the Banach space consisting of all continuous mappings  $Y: [0, T] \rightarrow L^2(\Omega, \mathcal{F}, \mathbb{P}; H)$  which are adapted to  $W$  (that is, such that for all  $t \in [0, T]$ ,  $Y(t)$  is  $\mathcal{F}_t$ -measurable, where  $\mathcal{F}_t$  is the  $\sigma$ -algebra generated by  $\{W(s), s \in [0, t]\}$ ), endowed with the norm,

$$\|Y\|_{C_W([0, T]; H)} = \left( \sup_{t \in [0, T]} \mathbb{E}(|Y(t)|^2) \right)^{1/2}.$$

It is called the space of all *mean square continuous adapted processes* on  $[0, T]$  taking values on  $H$ .

Equation (17) can be solved easily by a standard fixed-point argument in the Banach space  $C_W([0, T]; H)$ . Thus we have the result

**Proposition 9** *Equation (17) has a unique solution  $X(\cdot, x)$ .*

One checks easily that the corresponding transition semigroup  $P_t$  defined by (7) is Feller. When  $B = I$  one can show that  $P_t$  is strongly Feller by using the *Bismut-Elworthy* formula [10,40] for the derivative of  $D_x P_t \varphi(x)$  which reads as follows:

$$\langle DP_t \varphi(x), h \rangle = \frac{1}{t} \mathbb{E} \left[ \varphi(X(t, x)) \int_0^t \langle D_x X(s, x) \cdot h, dW(s) \rangle \right].$$

In this case one can also show that  $P_t$  is irreducible [20]. So, in view of Doob's theorem there is at most one invariant measure.

Existence of an invariant measure can be proved under the assumption

$$\langle Ax, x \rangle + \langle b(x), x \rangle \leq a - b|x|^2, \quad \forall x \in D(A),$$

where  $a, b > 0$ . In fact in this case one finds easily by Itô's formula an estimate of the second moment of  $X(t, x)$  independent of  $t$  of the form

$$\mathbb{E}|X(t, x)|^2 \leq c(1 + |x|^2). \quad (18)$$

By (18) the tightness of  $(\pi_t(x, \cdot))_{t \geq 0}$  follows for any fixed  $x \in H$ . We have in fact for any  $R > 0$

$$\begin{aligned} \pi_t(x, B_R^c) &\leq \frac{1}{R^2} \int_H |y|^2 \pi_t(x, dy) \\ &= \frac{1}{R^2} \mathbb{E}|X(t, x)|^2 \leq \frac{c}{R^2} (1 + |x|^2), \end{aligned}$$

where  $B_R^c$  is the complement of the ball  $B_R$  of center 0 and radius  $R$ . So, the existence of an invariant measure follows from the Krylov-Bogoliubov theorem.

### Reaction-Diffusion Equations Perturbed by Noise

We consider here a stochastic perturbation of problem (4) from Example 2, but for the sake of simplicity we take  $\mathcal{O} = [0, 1]$  and  $B = I$ . So, we have  $H = L^2(0, 1)$  and (4) becomes

$$\begin{cases} dX(t, \xi) = [D_\xi^2 X(t, \xi) + p(X(t, \xi))] dt + dW(t, \xi), \\ \quad \xi \in [0, 1], \\ X(t, 0) = X(t, 1) = 0, \quad t \geq 0, \\ X(0, \xi) = x(\xi), \quad \xi \in [0, 1], \quad x \in H, \end{cases} \quad (19)$$

where  $p$  is a polynomial of degree  $d > 1$  with negative leading coefficient.

For generalizations to systems of reaction-diffusion equations in bounded subsets of  $\mathbb{R}^n$  and for equations with multiplicative noise see [15,16] and references therein. For problems with reflexion see [65,82].

Now we write (19) in the mild form

$$X(t) = e^{tA}x + \int_0^t e^{(t-s)A}p(X(s))ds + W_A(t), \quad t \geq 0, \quad (20)$$

where  $W_A(t)$  is the stochastic convolution defined by (14). Moreover,

$$Ax = D_\xi^2 x, \quad x \in D(A) = H^2(0, 1) \cap H_0^1(0, 1)$$

and

$$b(x)(\xi) = p(x(\xi)), \quad x \in D(b) = L^{2d}(0, 1).$$

We notice that, since the leading coefficient of  $p$  is negative,  $b$  enjoys the basic dissipativity property

$$\langle b(x) - b(y), x - y \rangle \leq c_1 |x - y|^2, \quad \forall x, y \in L^{2d}(0, 1),$$

where

$$c_1 = \sup_{\xi \in \mathbb{R}} p'(\xi).$$

The operator  $A$  is self-adjoint and possesses a complete orthonormal system  $(e_k)$  of eigenfunctions, namely,

$$e_k(\xi) = (2/\pi)^{1/2} \sin(\pi\xi), \quad \xi \in [0, 1], \quad k \in \mathbb{N}.$$

Moreover

$$Ae_k = -\pi^2 k^2 e_k, \quad k \in \mathbb{N}.$$

Therefore, we have  $\langle Ax, x \rangle \leq -\pi^2 |x|^2$ , for all  $x \in H$ .

Let us give the definition of the mild solution of (19). We say that  $X \in C_W([0, T]; H)$  ( $C_W([0, T]; H)$  was defined in Sect. "Smooth Perturbations of System (12)" is a mild solution of problem (19) if  $X(t) \in L^{2d}(0, 1)$  for all  $t \geq 0$  and fulfills (20).

We notice that the condition  $X(t) \in L^{2d}(0, 1)$  is necessary in order for the integrand in (20) be meaningful.

The basic existence and uniqueness result is the following [3,26,41].

**Theorem 10** For any  $x \in L^{2d}(\mathcal{O})$ , there exists a unique mild solution  $X(\cdot, x)$  of problem (19).

*Proof* First we consider a regular approximating problem,

$$\begin{cases} dX_\alpha(t) = (AX_\alpha(t) + b_\alpha(X_\alpha(t))dt + dW(t), \\ X_\alpha(0) = x \in H, \end{cases} \quad (21)$$

where for any  $\alpha > 0$ ,  $b_\alpha$  is of class  $C^1$ ,  $b_\alpha(x) \rightarrow p(x)$  for all  $x \in L^{2d}(0, 1)$  as  $\alpha \rightarrow 0$  and the basic dissipativity property,

$$\langle b_\alpha(x) - b_\alpha(y), x - y \rangle \leq c_2 |x - y|^2, \quad \forall x, y \in H, \quad \alpha > 0,$$

is fulfilled, where  $c_2$  is a constant independent of  $\alpha$ . By Proposition 9 we know that problem (21) has a unique mild solution  $X_\alpha$ , that is,

$$X_\alpha(t) = e^{tA}x + \int_0^t e^{(t-s)A}b_\alpha(X(s))ds + W_A(t), \quad \mathbb{P}\text{-a.s.}$$

Taking advantage of the dissipativity of  $b_\alpha$  it is possible to find suitable estimates for  $X_\alpha$  and to show that  $X_\alpha$  converge, as  $\alpha \rightarrow 0$ , to the unique solution  $X$  of (19).  $\square$

By Theorem 10 we can construct a transition semigroup  $P_t$  on  $B_b(L^{2d}(0, 1))$ . However, it is useful to extend this semigroup to  $B_b(L^2(0, 1))$ . For this we need a weaker notion of solution of (19) when  $x \in H = L^2(0, 1)$ .

Let  $x \in H$ . We say that  $X \in C_W([0, T]; H)$  is a generalized solution of problem (19) if there exists a sequence  $(x_n) \subset L^{2d}(\mathcal{O})$ , such that

$$\lim_{n \rightarrow \infty} x_n = x \quad \text{in } H,$$

and

$$\lim_{n \rightarrow \infty} X(\cdot, x_n) = X(\cdot, x) \quad \text{in } C_W([0, T]; H).$$

It is easy to see, using again dissipativity of  $p$ , that this definition does not depend on the choice of the sequence  $(x_n)$  and to prove the following result.

**Corollary 11** For any  $x \in H$ , there exists a unique generalized solution  $X(\cdot, x)$  of problem (19)

Now we can define the transition semigroup

$$P_t \varphi(x) = \mathbb{E}[\varphi(X(t, x))],$$

for all  $\varphi \in B_b(H)$ .

We finally discuss invariant measures of  $P_t$  assuming for simplicity that  $p'(\xi) \leq 0$  so that  $p$  is dissipative. In this case we can show the following result [26].

**Theorem 12** The semigroup  $P_t$  has a unique invariant measure  $\nu$ , which is ergodic and strongly mixing and,

$$\lim_{t \rightarrow +\infty} P_t \varphi(x) = \int_H \varphi(y) \nu(dy), \quad x \in H. \quad (22)$$

*Proof* Let us consider the reaction-diffusion equation starting from  $-s$  where  $s > 0$ ,

$$\begin{cases} dX(t, \xi) = [D_\xi^2 X(t, \xi) + p(X(t, \xi))] dt + dW(t, \xi), \\ \xi \in [0, 1], \\ X(t, 0) = X(t, 1) = 0, \quad t \geq 0, \\ X(-s, \xi) = x(\xi), \quad \xi \in [0, 1], \quad x \in H. \end{cases} \quad (23)$$

Obviously we shall extend  $W(t)$  to negative times setting

$$W(t) = W(-t) \quad \text{if } t \leq 0.$$

Let us denote by  $X(t, -s, x)$  the solution of (23). Then, using the dissipativity of  $p$ , one can show that there exists  $\zeta \in L^2(\Omega, \mathcal{F}, \mathbb{P}; H)$  such that

$$\lim_{s \rightarrow +\infty} X(t, -s, x) = \zeta \quad \text{in } L^2(\Omega, \mathcal{F}, \mathbb{P}; H), \quad x \in H,$$

where  $\zeta$  is independent of  $t$ . Now, using the fact that the law of  $X(t, -s, x)$  coincides with that of  $X(t + s, x)$ , it follows that the law  $\nu$  of  $\zeta$  is an invariant measure for  $P_t$ .

Moreover, one can show that  $P_t$  is irreducible and strongly Feller, so the uniqueness of  $\nu$  as well as (22) follow from Doob's theorem.  $\square$

### Burgers Equations Perturbed by Noise

We are concerned with the Burgers equation in  $[0, 2\pi]$  perturbed by white noise

$$\begin{cases} dX(t, \xi) = \left( \partial_{\xi}^2 X(t, \xi) + \frac{1}{2} \partial_{\xi} (X^2(t, \xi)) \right) dt \\ \quad + dW(t, \xi), \quad \xi \in [0, 2\pi], \quad t > 0, \\ X(t, 0) = X(t, 2\pi), \quad t > 0 \\ X(0, \xi) = x(\xi), \quad \xi \in [0, 2\pi]. \end{cases} \quad (24)$$

We set  $H = L^2(0, 2\pi)$  and give the following definition. A *mild solution* in  $[0, T]$  of (24) is a process  $X \in C_W([0, T]; H)$  such that

$$X(t) = e^{tA}x + \frac{1}{2} \int_0^t \partial_{\xi} e^{(t-s)A} (X^2(s)) ds + W_A(t), \quad t \geq 0, \quad (25)$$

where the stochastic convolution  $W_A(t)$  is defined by (14).

We note that (25) is meaningful thanks to the estimate

$$\left| \partial_{\xi} e^{(t-s)A} y \right| \leq \kappa t^{-\frac{3}{4}} \|y\|_1, \quad x \in L^1(0, 2\pi)$$

(which can be proved by an elementary argument of harmonic analysis).

Now we state, following [29], an existence and uniqueness result.

**Theorem 13** *For any  $x \in H$  and  $T > 0$  there exists a unique mild solution  $X \in C_W([0, T]; H)$  of (24). Moreover, there exists a unique invariant measure which is ergodic and strongly mixing.*

*Proof* By the contraction principle it is easy to show existence and uniqueness of a local solution of (25) in a stochastic interval  $[0, \tau(\omega))$  (since (25) is a semilinear equation with a locally Lipschitz nonlinearity). In order to show global existence one needs to find an a priori estimate for  $\mathbb{P}$ -almost all  $\omega \in \Omega$ . For this we first reduce (25) to a deterministic equation (more precisely to a family of deterministic equations indexed by  $\omega \in \Omega$ ), setting

$$Y(t) = X(t) - W_A(t).$$

$Y(t)$  fulfills in fact the equation

$$\begin{cases} \frac{dY(t)}{dt} = AY(t) + b(Y(t) + W_A(t)), \quad t \in [0, T], \\ Y(0) = x. \end{cases}$$

An a priori estimate for  $Y(t)$  can be found by some manipulations using the basic property of  $b$ ,

$$\langle b(x), x \rangle = 0, \quad \forall x \in D(b).$$

For details and for the existence of an invariant measure see [29], for the uniqueness see [27].  $\square$

Another approach for studying the Burgers equation can be found in [9].

### 2D Navier–Stokes Equation Perturbed by Noise

We are here concerned with the equation

$$\begin{cases} dZ = (\Delta_{\xi} Z - Z \cdot \nabla Z + \nabla p) dt + B_1 dW(t) \\ \quad \text{in } [0, +\infty) \times \mathcal{O}, \\ \operatorname{div} Z = 0 \quad \text{in } [0, +\infty) \times \mathcal{O}, \\ Z(t, \cdot) \text{ is periodic with period } 2\pi, \\ Z(0, \cdot) = z \quad \text{in } \mathcal{O}, \end{cases} \quad (26)$$

where  $Z = (Z_1, Z_2)$  belongs to the space  $H$  of all square integrable divergence free vectors and  $B_1 \in L(H)$ .

Several papers have been devoted to stochastic 2D Navier–Stokes equations. We mention, besides the pioneering work [1,7,13,22]. Concerning 3D Navier–Stokes equations (which we do not consider here) one can look at [13,14,23,42,43,44,46,47].

We write (26) in the mild form

$$X(t) = e^{tA}x + \int_0^t e^{(t-s)A} b(X(s)) ds + W_A(t), \quad (27)$$

where the operators  $A$  and  $b$  were defined in Example 4 and the stochastic convolution  $W_A(t)$  is given by (14) with  $B = \mathcal{P}B_1$ .

In order to study (27) it is convenient to introduce the mapping

$$\Gamma(f) = \int_0^t e^{(t-s)A} b(f(s)) ds, \quad f \in L^4((0, T) \times \mathcal{O}), \quad t \geq 0,$$

and then to write (27) as

$$X(t) = e^{tA}x + \Gamma(X)(t) + W_A(t).$$

In fact one can show [27] that  $\Gamma$  maps  $L^4((0, T) \times \mathcal{O})$  into itself and for any  $f, g \in L^4((0, T) \times \mathcal{O}) := L^4$  we have

$$\|\Gamma(f) - \Gamma(g)\|_{L^4} \leq 4(\|f\|_{L^4} + \|g\|_{L^4}) \|f - g\|_{L^4}.$$

**Theorem 14** *For any  $x \in H$  there exists a unique mild solution  $X(\cdot, x)$  of (26).*

*Proof* One reduces (27) to a family of deterministic equations as in the case of the Burgers equation. Then one uses a fixed-point argument in the space  $L^4(0, T; L^4)$  [27].  $\square$

Existence of invariant measures is not difficult. It can be obtained, as in the previous examples, by the Krylov–Bogoliubov theorem. Uniqueness is more delicate. When the noise  $BdW(t)$  is not too degenerate it was proved in [45]. Otherwise one has to use the *coupling* argument [54,58,60,81].

### Future Directions

The number of papers devoted to nonlinear SPDEs is increasing. In fact, several models arising in the application are more realistic when a random perturbation is taken into account. Among new directions of research we mention the following.

- The direct study of the Kolmogorov equation (8) (considered as a parabolic equation with infinitely many variables). This can give important information about the properties of system (2) even in the case when the problem is not well posed. A result in this direction can be found in [28]. For a solution of (8) in the case of the 3D Navier–Stokes equation (with a failed attempt to prove uniqueness in law) see [24].
- Hamilton–Jacobi equations such as

$$\begin{cases} \frac{du(t, x)}{dt} = K_0 u(t, x) + \mathcal{H}(u_x(t, x)), \\ u(0, x) = \varphi(x), \end{cases} \quad x \in H, \quad t \geq 0,$$

where  $\mathcal{H}$  is a Hamiltonian. This is an important subject related to stochastic optimal control problems for SPDEs. For early results see Chap. 12 in [28]. Recent results are also connected with backward equations. Here the pioneering finite-dimensional result in [67] has been extended in infinite dimensions [50].

We believe also that an extension of the following topics to infinite dimensions would be interesting:

- Homogenization and averaging. In finite dimension see, e. g., [8,49]
- Measures solutions of Kolmogorov equations. In finite dimension this is a well-known subject, see, e. g., [11,12].

### Bibliography

- Albeverio S, Cruzeiro AB (1990) Global flows with invariant (Gibbs) measures for Euler and Navier–Stokes two dimensional fluids. *Commun Math Phys* 129:431–444
- Albeverio S, Röckner M (1991) Stochastic differential equations in infinite dimensions: solutions via Dirichlet forms. *Probab Th Rel Fields* 89:347–386
- Bally V, Gyongy I, Pardoux E (1994) White noise driven parabolic SPDEs with measurable drift. *J Funct Anal* 120(2):484–510
- Barbu V, Bogachev VI, Da Prato G, Röckner M (2006) Weak solution to the stochastic porous medium equations: the degenerate case. *J Funct Anal* 235(2):430–448
- Barbu V, Da Prato G (2002) The stochastic nonlinear damped wave equation. *Appl Math Optimiz* 46:125–141
- Barbu V, Da Prato G, Röckner M (2008) Existence and uniqueness of nonnegative solutions to the stochastic porous media equation. preprint S.N.S. Indiana University Math J 57(1):187–212
- Bensoussan A, Temam R (1973) Équations stochastiques du type Navier–Stokes. *J Funct Anal* 13:195–222
- Bensoussan A, Lions JL, Papanicolaou G (1978) Asymptotic analysis for periodic structures. In: *Studies in mathematics and its applications*, vol 5. North-Holland, Amsterdam
- Bertini L, Cancrini N, Jona-Lasinio G (1994) The stochastic Burgers equation. *Comm Math Phys* 165(2):211–232
- Bismut JM (1984) Large deviations and the Malliavin calculus. *Progress in Mathematics* 45. Birkhäuser, Boston
- Bogachev VI, Da Prato G, Röckner M (2004) Existence of solutions to weak parabolic equations for measures. *Proc London Math Soc* 3(88):753–774
- Bogachev VI, Krylov NV, Röckner M (2006) Elliptic equations for measures: regularity and global bounds of densities. *J Math Pures Appl* 9(6):743–757
- Brzezniak Z, Capinski M, Flandoli F (1991) Stochastic partial differential equations and turbulence. *Math Models Methods Appl Sci* 1(1):41–59
- Capinski M, Cutland N (1994) Statistical solutions of stochastic Navier–Stokes equations. *Indiana Univ Math J* 43(3):927–940
- Cerrai S (2001) Second order PDE’s in finite and infinite dimensions: A probabilistic approach. In: *Lecture Notes in Mathematics*, vol 1762. Springer, Berlin
- Cerrai S (2003) Stochastic reaction-diffusion systems with multiplicative noise and non-Lipschitz reaction term. *Probab Th Rel Fields* 125:271–304
- Cerrai S, Freidlin M (2006) On the Smoluchowski–Kramers approximation for a system with an infinite number of degrees of freedom. *Probab Th Rel Fields* 135(3):363–394
- Crauel H, Debussche A, Flandoli F (1997) Random attractors. *J Dynam Differen Equ* 9:307–341
- Dalang R, Frangos N (1998) The stochastic wave equation in two spatial dimensions. *Ann Probab* 26(1):187–212
- Da Prato G (2004) Kolmogorov equations for stochastic PDEs. Birkhäuser, Basel
- Da Prato G, Debussche A (1996) Stochastic Cahn–Hilliard equation. *Nonlinear Anal* 26(2):241–263
- Da Prato G, Debussche A (2002) 2D Navier–Stokes equations driven by a space-time white noise. *J Funct Anal* 196(1):180–210
- Da Prato G, Debussche A (2003) Strong solutions to the stochastic quantization equations. *Ann Probab* 31(4):1900–1916
- Da Prato G, Debussche A (2003) Ergodicity for the 3D stochastic Navier–Stokes equations. *J Math Pures Appl* 82:877–947
- Da Prato G, Tubaro L (2000) A new method to prove self-adjointness of some infinite dimensional Dirichlet operator. *Probab Th Rel Fields* 118(1):131–145



26. Da Prato G, Zabczyk J (1992) Stochastic equations in infinite dimensions. Cambridge University Press, Cambridge
27. Da Prato G, Zabczyk J (1996) Ergodicity for infinite dimensional systems. In: London Mathematical Society Lecture Notes, vol 229. Cambridge University Press, Cambridge
28. Da Prato G, Zabczyk J (2002) Second order partial differential equations in Hilbert spaces. In: London Mathematical Society Lecture Notes, vol 293. Cambridge University Press
29. Da Prato G, Debussche A, Temam R (1994) Stochastic Burgers equation. *Nonlinear Differ Equ Appl* 4:389–402
30. Da Prato G, Röckner M, Rozovskii BL, Wang FY (2006) Strong solutions of stochastic generalized porous media equations: Existence, uniqueness and ergodicity. *Comm Partial Differ Equ* 31(1–3):277–291
31. Dawson DA (1975) Stochastic evolution equations and related measures processes. *J Multivariate Anal* 5:1–52
32. de Bouard A, Debussche A (1999) A stochastic nonlinear Schrödinger equation with multiplicative noise. *Comm Math Phys* 205(1):161–181
33. de Bouard A, Debussche A (2002) On the effect of a noise on the solutions of the focusing supercritical nonlinear Schrödinger equation. *Probab Th Rel Fields* 123(1):76–79
34. de Bouard A, Debussche A (2003) The stochastic nonlinear Schrödinger equation in  $H^1$ . *Stoch Anal Appl* 21(1):97–126
35. de Bouard A, Debussche A (2005) Blow-up for the stochastic nonlinear Schrödinger equation with multiplicative noise. *Ann Probab* 33(3):1078–1110
36. de Bouard A, Debussche A, Tsutsumi Y (2004–05) Periodic solutions of the Korteweg–de Vries equation driven by white noise (electronic). *SIAM J Math Anal* 36(3):815–855
37. Debussche A, Printems J (2001) Effect of a localized random forcing term on the Korteweg–de Vries equation. *J Comput Anal Appl* 3(3):183–206
38. Debussche A, Zambotti L (2007) Conservative stochastic Cahn–Hilliard equation with reflection. *Ann Probab* 35(5):1706–1739
39. Eckmann JP, Ruelle D (1985) *Rev Mod Phys* 53:643–653
40. Elworthy KD (1992) Stochastic flows on Riemannian manifolds. In: Pinsky MA, Wihstutz V (eds) *Diffusion processes and related problems in analysis*, vol II (Charlotte, NC, 1990), *Progr Probab* 27. Birkhäuser, Boston, pp 37–72
41. Faris WG, Jona-Lasinio G (1982) Large fluctuations for a nonlinear heat equation with noise. *J Phys A*:15:3025–3055
42. Flandoli F (1994) Dissipativity and invariant measures for stochastic Navier–Stokes equations. *Nonlinear Differ Equ Appl* 1:403–423
43. Flandoli F (1997) Irreducibility of the 3D stochastic Navier–Stokes equation. *J Funct Anal* 149:160–177
44. Flandoli F, Gątarek D (1995) Martingale and stationary solutions for stochastic Navier–Stokes equations. *Probab Th Rel Fields* 102:367–391
45. Flandoli F, Maslowski B (1995) Ergodicity of the 2-D Navier–Stokes equation under random perturbations. *Commun Math Phys* 171:119–141
46. Flandoli F, Romito M (2002) Partial regularity for the stochastic Navier–Stokes equations. *Trans Am Math Soc* 354(6):2207–2241
47. Flandoli F, Romito M (2006) Markov selections and their regularity for the three-dimensional stochastic Navier–Stokes equations. *C R Math Acad Sci Paris* 343(1):47–50
48. Fleming WH (1975) A selection-migration model in population genetics. *J Math Biol* 2(3):219–233
49. Freidlin M (1996) Markov processes and differential equations: asymptotic problems. In: *Lectures in Mathematics ETH Zürich*. Birkhäuser, Basel
50. Fuhrman M, Tessitore G (2002) Nonlinear Kolmogorov equations in infinite dimensional spaces: the backward stochastic differential equations approach and applications to optimal control. *Ann Probab* 30(3):1397–1465
51. Funaki T (1991) The reversible measures of multi-dimensional Ginzburg–Landau type continuum model. *Osaka J Math* 28(3):463–494
52. Funaki T, Olla S (2001) Fluctuations for  $\nabla\phi$  interface model on a wall. *Stoch Process Appl* 94(1):1–27
53. Funaki T, Spohn H (1997) Motion by mean curvature from the Ginzburg–Landau  $\nabla\phi$  interface model. *Comm Math Phys* 185(1):1–36
54. Hairer M, Mattingly JC (2006) Ergodicity of the 2D Navier–Stokes equations with degenerate stochastic forcing. *Ann Math* 3:993–1032
55. Jona-Lasinio G, Mitter PK (1985) On the stochastic quantization of field theory. *Commun Math Phys* 101(3):409–436
56. Krylov NV (1999) An analytic approach to SPDEs. In: *Stochastic partial differential equations: six perspectives*, *Math Surveys Monogr*, vol 64. Amer Math Soc, Providence, pp 185–242
57. Krylov NV, Rozovskii BL (1981) Stochastic evolution equations, Translated from *Itogi Naukii Tekhniki*, *Seriya Sovremennyye Problemy Matematiki* 14:71–146. *J Soviet Math* 14:1233–1277
58. Kuksin S, Shirikyan A (2001) A coupling approach to randomly forced randomly forced PDE's I. *Commun Math Phys* 221:351–366
59. Liskevich V, Röckner M (1998) Strong uniqueness for a class of infinite dimensional Dirichlet operators and application to stochastic quantization. *Ann Scuola Norm Sup Pisa Cl Sci* 4(XXVII):69–91
60. Mattingly J (2002) Exponential convergence for the stochastically forced Navier–Stokes equations and other partially dissipative dynamics. *Commun Math Phys* 230(3):421–462
61. Mikulevicius R, Rozovskii B (1998) Martingale problems for stochastic PDE's. In: Carmona RA, Rozovskii B (eds) *Stochastic partial differential equations: Six perspectives*. *Mathematical Surveys and Monograph*, vol 64. American Mathematical Society, pp 243–325
62. Millet A, Morien PL (2001) On a nonlinear stochastic wave equation in the plane: existence and uniqueness of the solution. *Ann Appl Probab* 11:922–951
63. Millet A, Sanz-Solé M (2000) Approximation and support theorem for a wave equation in two space dimensions. *Bernoulli* 6:887–915
64. Musiela M, Rutkowski M (1997) Martingale methods in financial modelling. In: *Applications of Mathematics (New York)*, vol 36. Springer, Berlin
65. Nualart D, Pardoux E (1992) White noise driven quasilinear SPDEs with reflection. *Prob Theory Rel Fields* 93:77–89
66. Pardoux E (1975) Équations aux dérivées partielles stochastiques nonlinéaires monotones. Thèse, Université Paris XI
67. Pardoux E, Peng SG (1990) Adapted solution of a backward stochastic differential equation. *Syst Control Lett* 14(1):55–61
68. Parisi G, Wu YS (1981) *Sci Sin* 24:483–490
69. Peszat S, Zabczyk J (2000) Nonlinear stochastic wave and heat equations. *Probab Th Rel Fields* 116(3):421–443

70. Petersen K (1983) Ergodic Theory. Cambridge, London
71. Prevot C, Röckner M (2007) A concise course on stochastic partial differential equations, Monograph 2006. In: Lecture Notes in Mathematics. Springer, Berlin
72. Priola E (1999) On a class of Markov type semigroups in spaces of uniformly continuous and bounded functions. *Studia Math* 136:271–295
73. Ren J, Röckner M, Wang FY (2007) Stochastic generalized porous media and fast diffusions equations. *J Diff Equ* 238(1):118–152
74. Rozovskii BL (2003) Linear theory and applications to non-linear filtering (mathematics and its applications). Kluwer, Dordrecht
75. Sanz-Solé M (2005) Malliavin calculus with applications to stochastic partial differential equations. In: Fundamental sciences. EPFL Press, Lausanne; distributed by CRC Press, Boca Raton
76. Simon B (1974) The  $P(\phi)_2$  Euclidean (quantum) field theory. Princeton University Press, Princeton
77. Stannat W (2000) On the validity of the log-Sobolev inequality for symmetric Fleming–Viot operators. *Ann Probab* 28(2):667–684
78. Temam R (1988) Infinite-dimensional dynamical systems in mechanics and physics. Springer, New York
79. Viot M (1976) Solution faibles d'équations aux dérivées partielles non linéaires. Thèse, Université Pierre et Marie Curie, Paris
80. Walsh JB (1986) An introduction to stochastic partial differential equations. In: *cole d't de probabilités de Saint-Flour, XIV—1984*. Lecture Notes in Math, vol 1180. Springer, Berlin, pp 265–439
81. Weinan E, Mattingly JC, Sinai YG (2001) Gibbsian dynamics and ergodicity for the stochastically forced Navier–Stokes equation. *Commun Math Phys* 224:83–106
82. Zambotti L (2001) A reflected stochastic heat equation as a symmetric dynamic with respect to 3-d Bessel Bridge. *J Funct Anal* 180(1):195–209

## Non-negative Matrices and Digraphs

ABRAHAM BERMAN<sup>1</sup>, NAOMI SHAKED-MONDERER<sup>2</sup>

<sup>1</sup> Technion – Israel Institute of Technology, Haifa, Israel

<sup>2</sup> Emek Yezreel College, Emek Yezreel, Israel

### Article Outline

Glossary

Definition of the Subject

Introduction

Matrices, Graphs and Digraphs

Irreducible Nonnegative Matrices and Their Digraphs

Special Irreducible Matrices

Reducible Nonnegative Matrices and Their Digraphs

Examples: Ranking Problems

The Laplacian Matrix of a Directed Graph

Research Directions

Bibliography

### Glossary

**Nonnegative matrix** A *nonnegative matrix* is a matrix whose entries are real nonnegative numbers. The matrix is *positive* if all its elements are positive.

**Digraph** A *digraph* (*directed graph*)  $\Gamma$  consists of a finite set  $V$  and a set  $E$  of ordered pairs of elements of  $V$ . The elements of  $V$  are called *vertices* and those of  $E$  are called *arcs*. It is often represented graphically by points as the vertices of  $\Gamma$ , and arrows between these points as the arcs.

**Irreducible matrix** A *reducible matrix* is a matrix which is either the  $1 \times 1$  zero matrix, or a square matrix that has a zero submatrix on complementary sets of rows and columns. A square matrix which is not reducible is *irreducible*.

**Strongly connected digraph** A *strongly connected digraph* is a digraph in which there is a walk from every vertex to every other vertex.

**(Directed) walk** A *directed walk* (*walk*) in a digraph is a sequence of arcs of the digraph such that each arc starts at the end vertex of its predecessor.

### Definition of the Subject

The interplay between Matrix Theory and Graph Theory is fascinating and fruitful, e.g. [4,5,8,14,18,21,25,31,32]. Here we only deal with the interplay between nonnegative matrices and digraphs.

Digraphs are of great importance in Computer Science, Social Sciences and Natural Sciences, see [1,2,55]. Nonnegative matrices have applications to a variety of disciplines, such as numerical analysis, probability, game theory, economics, optimization, dynamical systems, and data mining, see [3,6,7,39]. These two important subjects, digraphs and nonnegative matrices, are also strongly connected. (Nonnegative) matrices are a natural way to describe digraphs and conversely, digraphs describe the zero-nonzero structure of a matrix. This survey concentrates on the meeting points of the two subjects.

The theory of nonnegative matrices is a hundred years old, and digraphs have accompanied it almost since its beginning. This continues to be true today, when modern applications stimulate the study of both subjects. One of the main sources for the renewed interest is probably the possibilities offered, and the problems generated, by the use of computers. We start our survey with an example of such application.

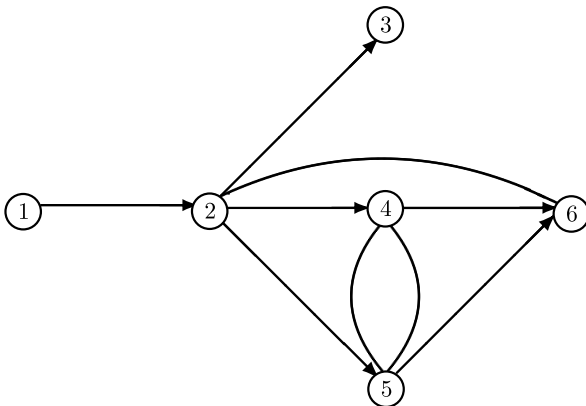
### Introduction

Google PageRank is a famous application of the theory of nonnegative matrices. Developed by Sergey Brin and

Lawrence Page [11], it ranks web pages according to their importance, under the principle that a page is important if important pages point to it. This is the idea: Suppose there are  $N$  web pages on the Web. Many of them contain links to other web pages. This can be described by a digraph: each web page is represented by a vertex, and a link on page  $i$  to page  $j$  is represented by an arc from  $i$  to  $j$ . The digraph in Fig. 1 shows a web of  $N = 6$  pages. (In the real Web, the number of pages is of course huge and constantly growing, with current estimates above  $N = 10^{10}$  [26]).

Let  $p_{ij}$  denote the probability that a surfer visiting page  $i$  will move to page  $j$ . Suppose there are  $d_i$  links out of page  $i$ . In a simplified version of the Google model, in the case that  $d_i > 0$ ,  $p_{ij} = \alpha \cdot \frac{1}{d_i} + (1 - \alpha) \cdot \frac{1}{N}$  if there is a link from page  $i$  to page  $j$ , and  $p_{ij} = (1 - \alpha) \cdot \frac{1}{N}$  if there is no such link, where  $\alpha = 0.85$ . For a page  $i$  with no outgoing links, that is, when  $d_i = 0$ ,  $p_{ij} = \frac{1}{N}$  for every  $j$ . This represents the idea that with probability  $\alpha$  a surfer on a page which has at least one link, randomly chooses (with equal probabilities) to follow one of the links on the page, and with probability  $1 - \alpha$  the surfer ignores the links, and chooses to surf to a random page on the Web (again, with equal probabilities), while if there are no links on the page, the surfer just chooses a random Web page to surf to. Note that the matrix  $P = [p_{ij}]$  has positive entries, and it is row stochastic (i. e., each of its rows sums up to 1). For the web described in Fig. 1,

$$P = \begin{bmatrix} 1/40 & 7/8 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 19/80 & 19/80 & 19/80 & 19/80 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/40 & 1/40 & 1/40 & 1/40 & 9/20 & 9/20 \\ 1/40 & 1/40 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}.$$



Non-negative Matrices and Digraphs, Figure 1  
A 6-pages web

The PageRank of a page  $i$ ,  $PR(i)$ , is defined as the probability that a random surfer will be in  $i$ . In this model, the larger  $PR(i)$  is, the more important page  $i$  is. The relations between these PageRanks are:

$$\begin{cases} PR(1) = p_{11}PR(1) + \cdots + p_{N1}PR(N) \\ \vdots \\ PR(N) = p_{1N}PR(1) + \cdots + p_{NN}PR(N) \end{cases}.$$

Let  $\pi = [PR(1), PR(2), \dots, PR(N)]^T$  be the (column) vector of PageRanks. By the above equations, the vector  $\pi$  satisfies  $\pi = P^T \pi$ . That is, the PageRank vector  $\pi$  is an eigenvector of  $P^T$  that corresponds to the eigenvalue 1, with nonnegative elements which sum up to 1. For the model to work we need to know that such an eigenvector exists, and that it is unique. This is guaranteed by the first major theorem on nonnegative matrices, which was proved by Perron in 1907 [43].

The Perron Theorem is a theorem on positive matrices. Its important extension to nonnegative matrices, proved by Frobenius in 1912 [23] and referred to as the Perron–Frobenius Theorem, requires the notion of irreducibility. This is the natural place to use graph theory but this was not done by Frobenius who did not care much for the, infant at his time, theory of graphs. In a very interesting article [48] Schneider describes the relations between irreducibility, the related concept of full indecomposability and the Frobenius–König Theorem as they appear in the papers of Frobenius, König and Markov (after which the highly important Markov Chains are named). He also touches the acrimonious controversy developed between Frobenius and König about their respective contributions to the topic. Without taking sides in this controversy, it is clear from historical perspective that Frobenius underestimated the relevance of digraphs to the theory of non-negative matrices. This hopefully is demonstrated in this survey.

### Matrices, Graphs and Digraphs

While we concentrate here on the interrelations between *nonnegative matrices* and *directed graphs*, we mention briefly the relations between *matrices* and *graphs*, and include the basic terminology of graphs and digraphs that will be used in this article.

A (*simple, undirected*) graph  $G$  is a pair  $G = (V, E)$ , where  $V$  is a finite set, and  $E$  is a set of unordered pairs of distinct elements in  $V$ . The elements of  $V$  are the *vertices* of  $G$ , and the elements of  $E$  are the *edges* of  $G$ . If  $e = \{u, v\}$  is an edge of  $G$ , we say that  $u$  (or  $v$ ) and  $e$  are *incident* with

each other, and that  $u$  and  $v$  are *adjacent* vertices in  $G$ . The *degree* of a vertex  $v \in V$  is the number of vertices adjacent to  $v$ . The added adjective ‘simple’ means that in these graphs there are no loops (that is, no edge joins a vertex to itself) or multiple edges (that is, any two distinct vertices are joined by at most one edge).

Several types of graphs may be associated with matrices. Given an  $m \times n$  matrix  $A$ , its *bipartite graph*  $B(A) = (V, E)$  has  $m + n$  vertices:  $V = \{r_1, \dots, r_m, c_1, \dots, c_n\}$  (each vertex  $r_i$  corresponds to a row of  $A$ , and each vertex  $c_j$  corresponds to a column of  $A$ ).  $\{u, v\}$  is an edge of  $B(A)$  if and only if  $u = r_i$  and  $v = c_j$  for some  $i$  and  $j$  and  $a_{ij} \neq 0$ .

With a *symmetric*  $n \times n$  matrix  $A$  we may associate its *graph*  $G(A) = (V, E)$ , with vertices  $V = \{1, \dots, n\}$  and  $\{i, j\} \in E$  if and only if  $i \neq j$  and  $a_{ij} (= a_{ji}) \neq 0$ .  $G(A)$  is a simple graph, and it conveys the zero-nonzero pattern of  $A$ , except for the diagonal entries of  $A$ .

We may also associate matrices with given graphs. Given a graph  $G = (V, E)$  with vertices  $V = \{v_1, \dots, v_n\}$  and edges  $E = \{e_1, \dots, e_m\}$ , its *incidence matrix*  $C = C(G)$  is the  $n \times m$  matrix with  $c_{ij} = 1$  if  $v_i$  is incident with  $e_j$  in  $G$ , and  $c_{ij} = 0$  otherwise. The *adjacency matrix* of  $G$ ,  $A = A(G)$  is the  $n \times n$   $(0, 1)$ -matrix with  $a_{ij} = 1$  if and only if  $\{v_i, v_j\} \in E$ . (The diagonal of  $A(G)$  is zero). Another matrix associated with  $G$  is its (combinatorial) *Laplacian*,  $L = L(G) = D(G) - A(G)$ , where  $D = D(G)$  is the diagonal matrix of vertex degrees, i.e.,  $d_{ii}$  is equal to the degree of the vertex  $v_i$ . By scaling  $L$  one obtains the (normalized) *Laplacian*  $\mathcal{L} = \mathcal{L}(G)$ , defined as  $\mathcal{L} = D^{-1/2} L D^{-1/2}$  (for this purpose  $D^{-1}$  is actually  $D^\dagger$ , the Moore–Penrose inverse of  $D$ : it is the diagonal matrix with  $(D^{-1})_{ii} = 0$  if  $d_{ii} = 0$ , and  $(D^{-1})_{ii} = 1/d_{ii}$  if  $d_{ii} > 0$ ). These matrices associated with the graph  $G$  are related:  $A = CC^T - D$ , and hence  $L = D - A = 2D - CC^T$ . The adjacency matrix  $A$  and the Laplacians  $L$  and  $\mathcal{L}$  are symmetric matrices, and thus have real eigenvalues. The relations between these eigenvalues and the structure of the graph  $G$  were extensively studied. In particular,  $L$  (and thus also  $\mathcal{L}$ ) is a singular positive semidefinite matrix, its smallest eigenvalue is zero ( $Le = 0$ , where  $e$  is the  $n$ -vector with all entries equal to 1). The second smallest eigenvalue of  $L$  was called by Fiedler the *algebraic connectivity* of  $G$ , motivated by the fact that the algebraic connectivity is nonzero if and only if the graph  $G$  is connected.

While there is a natural relation between symmetric matrices and graphs, general square matrices may be associated with digraphs. A *digraph* (directed graph)  $\Gamma$  is a pair  $\Gamma = (V, E)$ , where  $V$  is a finite set (we usually assume  $V = \{1, \dots, n\}$ ) and  $E$  is a set of ordered pairs of

(not necessarily distinct) elements of  $V$ . The elements of  $V$  are called the *vertices* of  $\Gamma$ ; The elements of  $E$  are the *arcs* of  $\Gamma$ . We say that  $(i, j) \in E$  is an arc *from*  $i$  *to*  $j$ ;  $i$  is its *initial vertex* or *tail*, and  $j$  is its *terminal vertex* or *head*. An arc  $(i, i)$  is called a *loop* ( $i$  is both its initial vertex and its terminal vertex). The number of arcs with  $i$  as an initial vertex is the *outdegree* of  $i$ . The number of arcs with  $i$  as terminal vertex is the *indegree* of  $i$ . We will consider only digraphs with no multiple arcs. That is, for every two distinct vertices  $i$  and  $j$  there is at most one arc from  $i$  to  $j$ . A digraph which has no loops and no multiple arcs is called *simple*. So in this article, a simple digraph is simply a digraph with no loops.

A *weighted digraph* is a digraph in which each arc  $(i, j)$  is assigned a weight  $\omega_{ij}$  (usually a real or complex number). A (*directed*) *walk* of length  $m$  from vertex  $i$  to vertex  $j$  in a digraph  $\Gamma$  is a sequence of  $m$  arcs  $(i_0, i_1), (i_1, i_2), \dots, (i_{m-1}, i_m)$ , with the initial vertex of each arc equal to the terminal vertex of its predecessor, and  $i_0 = i, i_m = j$ . If  $i_0 = i_m$ , it is a *closed walk*. If all the vertices in the walk are distinct, except, possibly, the first and last, the walk is called a (*directed*) *path*. When  $i_m = i_0$  the path is called a (*directed*) *cycle*. The *distance* from vertex  $i$  to vertex  $j$  is the length of the shortest path from  $i$  to  $j$ , and the *diameter* of a digraph is largest distance between two vertices in the digraph. The *girth* of a digraph is the length of its shortest cycle.

A digraph is *strongly connected* if for every two vertices  $i \neq j$  there are walks from  $i$  to  $j$  and from  $j$  to  $i$ . Every digraph on one vertex is strongly connected.

With each simple digraph  $\Gamma$  there is an *underlying graph*. The underlying graph  $G$  has the same vertices as  $\Gamma$ , and each arc  $(i, j)$  of  $\Gamma$  is replaced by the corresponding edge  $\{i, j\}$  in  $G$ . (No multiple edges, though, so if  $(i, j)$  and  $(j, i)$  are both arcs of  $\Gamma$ , there is only one edge  $\{i, j\}$  in  $G$ ).

Let  $\Gamma$  be a digraph with vertices  $V = \{1, \dots, n\}$  and arcs  $E = \{e_1, \dots, e_m\}$ . Its *incidence matrix*  $Q = Q(\Gamma)$  is the  $n \times m$  matrix with  $q_{ij} = -1$  if  $i$  the initial vertex of  $e_j$ ,  $q_{ij} = 1$  if  $i$  the terminal vertex of  $e_j$  and  $q_{ij} = 0$  otherwise. (If  $\Gamma$  is a simple digraph with incidence matrix  $Q$ , and  $G$  is its underlying digraph with combinatorial Laplacian  $L$ , then  $QQ^T = L$ ).

The *adjacency matrix*  $A = A(\Gamma)$  of a digraph  $\Gamma$  with vertices  $\{1, \dots, n\}$  is an  $n \times n$  matrix with  $a_{ij} = 1$  if and only if  $(i, j) \in E$ , and  $a_{ij} = 0$  otherwise. The adjacency matrix is a square  $(0, 1)$ -matrix. On the other hand, given an  $n \times n$  matrix  $A$ , the *digraph* of  $A$ ,  $\Gamma(A)$ , has vertices  $\{1, \dots, n\}$ , and  $(i, j)$  is an arc if and only if  $a_{ij} \neq 0$ . We may also consider a *weighted digraph* associated with  $A$  by assigning the weight  $a_{ij}$  to the arc  $(i, j)$  of  $\Gamma(A)$ . Obviously,  $\Gamma(A(\Gamma)) = \Gamma$ .



Re-labeling the vertices of a digraph is equivalent to simultaneous permutation of the rows and the columns of an associated matrix. Such simultaneous permutation is achieved by pre-multiplying  $A$  by a permutation matrix  $P$ , and post-multiplying it by  $P^T$ . The digraph  $\Gamma(PAP^T)$  is obtained from the digraph  $\Gamma(A)$  by relabeling its vertices according to the permutation, and vice versa: if  $\Gamma'$  is a digraph obtained from  $\Gamma(A)$  by re-labeling the vertices, then there exists a permutation matrix  $P$  such that  $\Gamma(PAP^T) = \Gamma'$ .

A *subdigraph* of  $\Gamma$  is a digraph  $\Gamma' = (V', E')$ , where  $V' \subseteq V$  and  $E' \subseteq E$ . Such  $\Gamma'$  is an *induced subdigraph* of  $\Gamma$  if  $E' = E \cap (V' \times V')$ , i.e.,  $E'$  consists of all the arcs in  $E$  whose initial and terminal vertices are in  $V'$ . The adjacency matrix of the subdigraph of  $\Gamma$  induced by  $V'$  is a principal submatrix of the adjacency matrix  $A = A(\Gamma)$  (that is, a submatrix of  $A$  lying on rows and columns with the same indices). We denote this submatrix by  $A[V']$ . The converse also holds. If  $\Gamma = \Gamma(A)$  for some matrix  $A$ , then the digraph of a principal submatrix of  $A$  is an induced subdigraph of  $\Gamma$ .

In what follows, all matrices under discussion will be square, unless specifically stated otherwise.

As mentioned before, we deal here only with nonnegative matrices and digraphs. For more on the connections between matrices and graphs, see e.g., Part II of [31], and the reference there. For the many connections between the spectrum of the Laplacian of a graph to graph structure parameters see [18].

### Irreducible Nonnegative Matrices and Their Digraphs

A matrix  $A$  is *nonnegative* (denoted  $A \geq 0$ ) if all its entries are nonnegative, and *positive* (denoted  $A > 0$ ) if all its entries are positive. If  $A$  and  $B$  are matrices of the same order, we write  $A \geq B$  if  $A - B \geq 0$  and  $A > B$  if  $A - B > 0$ .

If  $A$  is an  $n \times n$  matrix, then its spectrum consists of  $n$  (complex, not necessarily distinct) eigenvalues,  $\lambda_1, \dots, \lambda_n$ . The *spectral radius* of  $A$  is  $\rho(A) = \max\{|\lambda_i| : 1 \leq i \leq n\}$ . Let  $\mu(A) = \max\{|\lambda_i| : 1 \leq i \leq n, \lambda_i \neq \rho(A)\}$ . With these notations, we can state Perron's Theorem [43]:

**Theorem 1 (Perron Theorem)** *If  $A$  is a positive square matrix, then*

- $\rho(A) > 0$ .
- $\rho(A)$  is a simple eigenvalue of  $A$  (called the Perron root of  $A$ ).
- To  $\rho(A)$  there corresponds a positive eigenvector. (Such an eigenvector, normalized so that the sum of its entries is 1, is called the Perron vector of  $A$ ).

- $\mu(A) < \rho(A)$ , that is,  $\rho(A)$  is the only eigenvalue of  $A$  of maximum modulus.
- $\lim_{m \rightarrow \infty} \left(\frac{A}{\rho(A)}\right)^m = L$ ,  $L = xy^T$ , where  $Ax = \rho(A)x$ ,  $x > 0$ ,  $A^T y = \rho(A)y$ ,  $y > 0$ , and  $x^T y = 1$ .
- For every  $\mu(A)/\rho(A) < r < 1$  there exists a constant  $C = C(r, A)$  such that for every  $m$ ,  $\left\|\left(\frac{A}{\rho(A)}\right)^m - L\right\|_\infty \leq Cr^m$ .

The Perron Theorem may be extended to a wider class of matrices. A matrix  $A$  is *primitive* if for some natural number  $m$ ,  $A^m$  is positive.

**Theorem 2** *In Perron's Theorem (Theorem 1) "positive" may be replaced by "primitive".*

The theorem can be generalized to irreducible matrices. An  $n \times n$  matrix  $A$  is *reducible* if  $n > 1$  and there exists a permutation matrix  $P$  such that

$$PAP^T = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix},$$

with square diagonal blocks  $B$  and  $D$ , or  $n = 1$  and  $A = 0$ . A matrix is *irreducible* if it is not reducible. Irreducibility may be described in terms of the digraph of  $A$ : A matrix  $A$  of order greater than 1 is irreducible if and only if  $\Gamma(A)$  is strongly connected. Moreover,

**Theorem 3** *Let  $A$  be an  $n \times n$  nonnegative matrix,  $n > 1$ . Then the following are equivalent:*

- $A$  is irreducible.
- $\Gamma(A)$  is strongly connected.
- There is a positive integer  $m$  such that  $(I + A)^m > 0$ .
- $(I + A)^{n-1} > 0$ .

In Perron's Theorem (Theorem 1) parts a–c, 'positive' may be replaced by 'nonnegative and irreducible' – see parts a–c in Theorem 4 below. The new parts d–f of Theorem 4 also hold for such matrices (in particular for positive matrices). Theorem 4, the Perron–Frobenius Theorem, was first proved in [23].

**Theorem 4 (Perron–Frobenius Theorem, part I)** *If  $A$  is an  $n \times n$  nonnegative and irreducible square matrix, then*

- $\rho(A) > 0$ .
- $\rho(A)$  is a simple eigenvalue of  $A$  (called the Perron root of  $A$ ).
- To  $\rho(A)$  there corresponds a positive eigenvector. (Such an eigenvector, normalized so that the sum of its entries is 1, is called the Perron vector of  $A$ ).

In addition,



- d. If  $Ax = \lambda x$  and the vector  $x$  is positive, then  $\lambda = \rho(A)$ .
- e. If  $B \geq A$  and  $B \neq A$ , then  $\rho(B) > \rho(A)$ .
- f. If  $0 \leq B \leq A$  and  $B \neq A$ , then  $\rho(B) < \rho(A)$ .

If a nonnegative  $A$  is irreducible, so is  $A^T$ . The spectrum of  $A^T$  is equal to that of  $A$ . Hence there exists a Perron vector  $y$  of  $A^T$ . It satisfies:  $y^T A = \rho(A)y^T$ . The Perron vectors  $x$  and  $y$  of  $A$  and  $A^T$  are referred to as the right- and left- Perron vectors of  $A$ , respectively.

The following result follows from Part f of the Theorem 4.

**Corollary 5** If  $B \neq A$  is a principal submatrix of a non-negative irreducible matrix  $A$ , then  $\rho(B) < \rho(A)$ .

When  $A$  is irreducible and nonnegative, but not primitive, there may be several eigenvalues of maximum modulus. The number of eigenvalues of  $A$  with modulus  $\rho(A)$  is called the *index of cyclicity* of  $A$ .

**Theorem 6 (Perron–Frobenius Theorem, part II)** If  $A$  is a nonnegative and irreducible square matrix with index of cyclicity  $k > 1$ , then

- a. The eigenvalues of  $A$  of modulus  $\rho(A)$  are  $\rho(A)e^{2\pi il/k}$ ,  $l = 0, 1, \dots, k-1$ .
- b. The spectrum of  $A$  goes onto itself under rotation of the complex plane by  $2\pi/k$ .
- c. There exists a permutation matrix  $P$  such that

$$PAP^T = \begin{bmatrix} 0 & A_{12} & 0 & \cdots & 0 \\ 0 & 0 & A_{23} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & A_{k-1,k} \\ A_{k1} & 0 & 0 & \cdots & 0 \end{bmatrix},$$

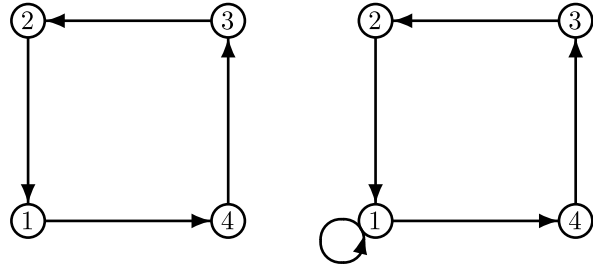
where the zero diagonal blocks are square.

**Corollary 7** A nonnegative irreducible matrix  $A$  is primitive if and only if its index of cyclicity is 1.

Using digraphs of nonnegative matrices simplifies the proof of the Perron–Frobenius Theorem. Digraphs are particularly useful in the computation of the index of cyclicity of an irreducible matrix [45]:

**Theorem 8 (Romanovsky Theorem)** Let  $A$  be an  $n \times n$  irreducible nonnegative matrix, and for every  $1 \leq i \leq n$  let  $d_i$  be the greatest common divisor of all the lengths of closed walks through  $i$  in  $\Gamma(A)$ . Then all the  $d_i$ 's are equal and are equal to the order of cyclicity of  $A$ .

An actual computation of the index of cyclicity by this theorem may be very hard when  $A$  is a large matrix and  $\Gamma(A)$  has many cycles, but can be done for small matrices:



**Non-negative Matrices and Digraphs, Figure 2**  
The digraphs of Examples 9

**Example 9** Let

$$A_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Then  $\Gamma(A_1)$  is the digraph shown on the left and  $\Gamma(A_2)$  is the digraph shown on the right in Fig. 2; both are strongly connected.

For  $\Gamma(A_1)$  we have  $d_1 = d_2 = d_3 = d_4 = 4$  and the index of cyclicity of  $A_1$  is 4. For  $\Gamma(A_2)$ , the existence of the loop at vertex 1 immediately implies that  $d_1 = d_2 = d_3 = d_4 = 1$  and the matrix is primitive.

Combining Romanovsky's Theorem and the Perron–Frobenius Theorem, part II, we get that

**Corollary 10** A nonnegative irreducible matrix  $A$  is primitive if  $\text{trace}(A) > 0$ .

Every primitive matrix is irreducible (see Theorem 3). However, the converse is not true. For example,

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

is irreducible and not primitive. Note that an  $n \times n$ ,  $n > 1$ , nonnegative matrix  $A$  is irreducible if and only if its digraph is strongly connected, and it is primitive if and only if there exists  $m$  such that for every pair of (not necessarily distinct) vertices  $u$  and  $v$  of  $\Gamma(A)$ , there is a walk of length  $m$  from  $u$  to  $v$ . The smallest such  $m$ , that is, the smallest  $m$  for which  $A^m$  is positive, is called the *exponent* of  $A$ . For information about exponents see Sect. 3.5 in [14]. The index of cyclicity is sometimes called the *period* of the matrix. A primitive matrix is sometimes called an *aperiodic* matrix. We mention also that in a different approach, every  $1 \times 1$  matrix is considered irreducible, even if it is zero. This corresponds well with the definition of every digraph on one vertex as strongly connected digraph, but requires some changes in the statement of Theorem 4.

The above results are classical, and details may be found in many books, e.g., Chap. 2 in [7], or Chap. 8 in [33].

## Special Irreducible Matrices

### Nearly Reducible Matrices

A strongly connected digraph  $\Gamma$  is *minimally connected* (or *minimally strong*) if removal of any arc of  $\Gamma$  results in a digraph which is not strongly connected. A matrix is *nearly reducible* if its digraph is minimally connected. That is, an  $n \times n$  matrix  $A$ , with  $n > 1$  is nearly reducible if it is irreducible, and replacing any one nonzero entry by a zero yields a reducible matrix. The  $1 \times 1$  zero matrix is the only  $1 \times 1$  nearly reducible matrix (though by our definition it is also reducible). A minimally connected digraph has no loops, and a nearly reducible matrix has zero diagonal.

Nearly reducible matrices have a specific zero-nonzero pattern, described inductively in the following theorem, due to Hartfiel [29]:

**Theorem 11** *Let  $A$  be an  $n \times n$  nearly reducible matrix,  $n \geq 2$ . Then there exists an integer  $1 \leq m < n$  and a permutation matrix  $P$  such that*

$$PAP^T = \begin{bmatrix} B & C \\ D & E \end{bmatrix},$$

where  $B$  is an  $m \times m$  matrix of the form

$$B = \begin{bmatrix} 0 & * & 0 & \cdots & 0 \\ 0 & 0 & * & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & * \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where each  $*$  stands for a nonzero entry,  $C$  is an  $m \times (n-m)$  matrix with a single nonzero entry, located in the last row,  $D$  is an  $(n-m) \times m$  matrix with a single nonzero entry, located in the first column, and  $E$  is an  $(n-m) \times (n-m)$  nearly reducible matrix.

The following theorem is a translation to nearly reducible matrices of a result of Gupta on minimally connected digraphs [27]:

**Theorem 12** *Let  $A$  be an  $n \times n$  nearly reducible matrix,  $n > 1$ . Then the number of nonzero entries in  $A$  is at least  $n$  and at most  $2(n-1)$ . The lower bound is attained when  $\Gamma(A)$  is an  $n$ -cycle, and only in this case. The upper bound is attained only when  $\Gamma(A)$  is a digraph obtained from a tree by replacing each edge  $\{i, j\}$  of the tree with two oppositely directed arcs,  $(i, j)$  and  $(j, i)$ .*

The results concerning nearly reducible matrices and minimally connected digraph may be found in Sect. 3.3 in [14], and Sect. 4.5 in [42]. See also Sect. 29.8 in [31].

### Fully Indecomposable Matrices

A concept related to irreducibility is that of full indecomposability. We should first mention that reducible and irreducible matrices are sometimes called *decomposable* and *indecomposable*, which explains the terminology of partly- and fully- decomposable matrices introduced below.

An  $n \times n$  matrix  $A$ , is *partly decomposable* if  $n > 1$  and  $A$  has a zero  $m \times k$  submatrix with  $m$  and  $k$  positive integers such that  $m + k = n$ , or  $n = 1$  and  $A = 0$ . That is, a square matrix  $A$  of order greater than 1 is partly decomposable if and only if there exist permutation matrices  $P$  and  $Q$  such that

$$PAQ = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix},$$

with square diagonal blocks  $B$  and  $D$ . A square  $n \times n$  matrix  $A$  is *fully indecomposable* if and only if it is not partly decomposable.

In terms of the bipartite graph of  $A$ :  $A$  is partly decomposable if there exists a subset  $S = \{r_{i_1}, \dots, r_{i_m}\}$  of the “row vertices”  $\{r_1, \dots, r_n\}$ , and a subset  $T = \{c_{j_1}, \dots, c_{j_{n-m}}\}$  of the “column vertices”  $\{c_1, \dots, c_n\}$ , such that there are no edges in  $B(A)$  from  $S$  to  $T$ . In terms of the digraph of  $A$ :  $A$  is partly decomposable if there exist subsets  $V_1$  and  $V_2$  of the vertex set  $\{1, \dots, n\}$ , not necessarily disjoint, of sizes  $m$  and  $k$  such that  $m + k = n$ , and such that there are no arcs in  $\Gamma(A)$  from  $V_1$  into  $V_2$ . In comparison, the matrix  $A$  is reducible if there exist such subsets  $V_1$  and  $V_2$  which are disjoint and  $V_1 \cup V_2 = \{1, \dots, n\}$ .

Clearly a reducible matrix of order greater than 1 is also partly decomposable, and a fully indecomposable matrix is irreducible. The converse does not hold: the matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

is an example of a partly decomposable matrix ( $V_1 = V_2 = \{1\}$ ) which is irreducible. However, irreducibility and full indecomposability are not too far apart, as the following results show. For the first of the two see Brualdi, Parter and Schneider [13] and for the second see Brualdi [12].

**Theorem 13** *A square matrix  $A$  of order greater than 1 is fully indecomposable if and only if for some permutation matrix  $P$  the matrix  $PA$  is irreducible with all the entries on its main diagonal non-zero.*

**Theorem 14** *A square nonnegative matrix  $A$  of order greater than 1 is irreducible if and only if  $I + A$  is fully indecomposable.*

Henrik Minc opened a course on nonnegative matrices that he gave at the Technion in 1974 (which was the basis for [42]) by stating that the Frobenius–König Theorem is the fundamental result on the zero pattern of square matrices.

**Theorem 15 (Frobenius–König Theorem)** *A necessary and sufficient condition that every diagonal of an  $n \times n$  matrix contains a zero is that the matrix contain an  $m \times k$  zero submatrix with  $m + k = n + 1$ .*

Recall that the *permanent* of a matrix  $A$  is given by

$$\text{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^n a_{i\sigma(i)},$$

where  $S_n$  is the group of all permutations on  $\{1, \dots, n\}$ . That is, it is the sum of all diagonal products of  $A$ .

**Corollary 16** *If  $A$  is a square nonnegative matrix and  $\text{per}(A) = 0$ , then  $A$  is partly decomposable.*

We mention two more results, due to Marcus and Minc [40]:

**Theorem 17** *A nonnegative fully indecomposable matrix is primitive.*

(The converse is not true:

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

is primitive but not fully indecomposable).

**Theorem 18** *A fully indecomposable  $n \times n$  matrix can contain at most  $n(n - 2)$  zeros.*

For details on fully indecomposable matrices see, e.g., Sect. II.5.4 in [41], and Sect. 4.2 in [14]. It is worth noting that the Frobenius–König Theorem played an important role in the proof of the Perron–Frobenius Theorem (see the discussion in [48]). Another important application of the Frobenius–König Theorem is Birkhoff’s Theorem [9], which states that the set of all  $n \times n$  doubly stochastic matrices, i.e. matrices that are both row stochastic and column stochastic, forms a convex polyhedron with the permutation matrices as vertices. A slightly more general form of the Frobenius–König Theorem is P. Hall’s theorem on systems of distinct representatives [28].

## Reducible Nonnegative Matrices and Their Digraphs

For reducible nonnegative matrices, there is also a strong connection between the spectral properties of the matrices and the graph theoretic properties of their digraphs. This connection has been studied extensively, and we review here a sample of the results.

If  $A$  is a reducible matrix, there is a permutation matrix  $P$  such that

$$PAP^T = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix}.$$

If any of the square diagonal blocks is reducible, we may further reduce the matrix  $A$ , until we get a lower block-triangular matrix, with irreducible and/or  $1 \times 1$  zero diagonal blocks, which is permutationally equivalent to  $A$ .

This can be also described by the digraph of  $A$ . Let  $\Gamma$  be any digraph on vertices  $V = \{1, \dots, n\}$ . We say that a vertex  $i$  has access to vertex  $j$  in  $\Gamma$  if  $i = j$  or there is a walk from  $i$  to  $j$  in  $\Gamma$ . We say that  $i$  and  $j$  communicate if  $i$  has access to  $j$  and  $j$  has access to  $i$ . Communication is an equivalence relation on  $V$ , and it therefore induces a partition of  $V$  into equivalence classes  $V_1, \dots, V_k$ . The induced digraphs  $\Gamma_i = \Gamma(V_i)$  are strongly connected. These are the *strongly connected components* of  $\Gamma$ .

Let  $R(\Gamma)$  be the digraph with vertices  $V_1, \dots, V_k$  and an arc from  $V_i$  to  $V_j$ ,  $i \neq j$ , if and only if there is a walk from a vertex in  $V_i$  to a vertex in  $V_j$  (and thus from any vertex in  $V_i$  to any vertex in  $V_j$ ).  $R(\Gamma)$  is called the *reduced graph* of  $\Gamma$ .

**Theorem 19** *Let  $\Gamma$  be a digraph. Then  $R(\Gamma)$  is a simple digraph with no cycles, and its vertices may be labeled  $V_1, \dots, V_k$  so that if  $(V_i, V_j)$  is an arc of  $R(\Gamma)$  then  $i > j$ .*

When  $\Gamma$  is  $\Gamma(A)$ , the digraph of a matrix  $A$ , the equivalence classes of the communication relation on  $\{1, \dots, n\}$  are called the *classes* of  $A$ . The reduced graph of  $\Gamma(A)$  is denoted also by  $R(A)$  and referred to as the *reduced graph* of  $A$ . The rows and columns of  $A$  may be permuted, according to the order of the classes described in Theorem 19. We thus get

**Theorem 20** *Let  $A$  be an  $n \times n$  matrix. Then there exists a permutation matrix  $P$  such that*

$$PAP^T = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{bmatrix}, \quad (1)$$

where each of the diagonal blocks  $A_{11}, \dots, A_{kk}$  is square and irreducible or a  $1 \times 1$  zero matrix. The diagonal blocks

are uniquely determined, up to a simultaneous permutation of rows and columns, but their ordering is not necessarily unique.

The block matrix (1) is the (lower block-triangular) Frobenius normal form of  $A$ . The spectrum of  $A$  is the union of the spectrums of the diagonal blocks in the Frobenius normal form. In particular, the spectral radius of  $A$ ,  $\rho(A)$ , is an eigenvalue of  $A$ , and there exists a class  $V_i$  of  $A$  such that  $\rho(A[V_i]) = \rho(A)$ . Such a class is called a *basic class*.

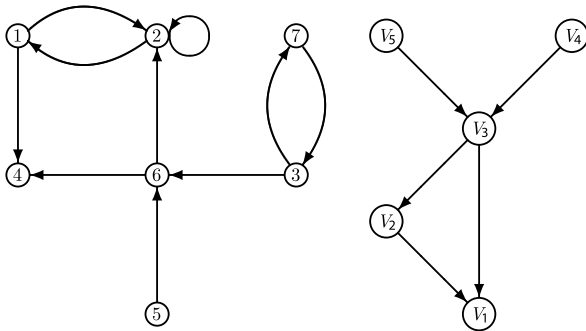
If there is a walk in  $R(A)$  from  $V_i$  to  $V_j$  or  $V_i = V_j$ , we say that the class  $V_i$  has access to the class  $V_j$ , and denote it by  $V_i \geq V_j$ . We will write  $V_i > -V_j$  if  $V_i$  has access to  $V_j$  and  $V_i \neq V_j$ . We will also write  $i > -V_j$  if  $i$  is a vertex and either  $i \in V_j$  or there is a walk from  $i$  to a vertex in  $V_j$ . A class is *final* if it has no access to another class (i.e., its outdegree in  $R(A)$  is 0). A class is *initial* if no other class has access to it (that is, its indegree in  $R(A)$  is 0). The *level* of a basic class  $V_i$  is the maximal number of basic classes on a path in  $R(A)$  that terminates at  $V_i$ . The level of an initial basic class is 1.

**Example 21** Let

$$A = \begin{bmatrix} 0 & 1 & 0 & 4 & 0 & 0 & 0 \\ 3 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 8 & 0 & 9 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then  $\Gamma(A)$  is the digraph shown on the left in Fig. 3, and  $R(A)$  is the digraph shown on the right.

The classes are  $\{1, 2\}$ ,  $\{3, 7\}$ ,  $\{4\}$ ,  $\{5\}$  and  $\{6\}$ . The classes  $\{5\}$  and  $\{3, 7\}$  are initial classes,  $\{4\}$  is a final class. One labeling that fits the requirements of Theorem 19 is  $V_1 = \{4\}$ ,  $V_2 = \{1, 2\}$ ,  $V_3 = \{6\}$ ,  $V_4 = \{5\}$  and



**Non-negative Matrices and Digraphs, Figure 3**  
The digraph and reduced digraph of Example 2

$V_5 = \{3, 7\}$ . Let

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

be the permutation matrix corresponding to the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 1 & 2 & 6 & 5 & 3 & 7 \end{pmatrix},$$

then

$$PAP^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 2 & 0 & 0 & 0 & 0 \\ 9 & 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \end{bmatrix}$$

is  $A$ 's Frobenius normal form. The spectra of the diagonal blocks are  $\{0\}$ ,  $\{3, -1\}$ ,  $\{0\}$ ,  $\{0\}$ ,  $\{3, -3\}$ . Hence  $V_2$  and  $V_5$  are the basic classes of  $A$ .

Finally, the basic classes  $V_2$  and  $V_5$  are of levels 2 and 1, respectively.

The algebraic multiplicity of the Perron eigenvalue  $\rho(A)$  is obviously equal to the number of basic classes of  $A$ . The size of the largest Jordan block corresponding to  $\rho(A)$  can also be "read" from the reduced graph of  $A$ . Recall that the size of the largest Jordan block associated with the eigenvalue  $\rho(A)$  is the multiplicity of  $\rho(A)$  as a root of the minimal polynomial of  $A$ . It is called the *index* of the Perron eigenvalue  $\rho(A)$  of the matrix  $A$  and we denote it by  $\nu = \nu(A)$ . Equivalently, the index of  $\rho(A)$  is the smallest nonnegative integer  $\nu$  such that  $\mathcal{N}((\rho(A)I - A)^\nu) = \mathcal{N}((\rho(A)I - A)^{\nu+1})$ , where  $\mathcal{N}(B)$  denotes the nullspace of a matrix  $B$ . The following was proved in [46]:

**Theorem 22 (Rothblum Index Theorem)** *Let  $A$  be a square nonnegative matrix. Then the maximal level of a basic class is equal to  $\nu(A)$ .*

The matrix  $A$  in Example 21 thus has index  $\nu(A) = 2$ .

Theorem 22 implies the following (earlier) result by Schneider [47]:

**Corollary 23** *Let  $A$  be a square nonnegative matrix. Then in the Jordan Canonical form of  $A$  there is only one*

block corresponding to the eigenvalue  $\rho(A)$  (equivalently,  $\dim \mathcal{N}((\rho(A)I - A)) = 1$ ) if and only if for every two basic classes in  $R(A)$ , one of the two classes has access to the other.

When  $A$  is a reducible nonnegative matrix, there need not exist a positive eigenvector corresponding to  $\rho(A)$ . The following theorem appeared in vol. II, p. 77 of [24]:

**Theorem 24** *Let  $A$  be a nonnegative matrix. There exists a positive eigenvector corresponding to the spectral radius of  $A$  if and only if the final classes of  $A$  are exactly the basic ones.*

The algebraic eigenspace (or the Perron generalized eigenspace) of  $A$ , denoted by  $E(A)$ , is defined to be  $\mathcal{N}((\rho(A)I - A)^n) = \mathcal{N}((\rho(A)I - A)^v)$ . The nonzero elements of the algebraic eigenspace of  $A$  are called *generalized eigenvectors* of  $A$  associated with the eigenvalue  $\rho(A)$ . There are several results on the existence of various nonnegative bases to the algebraic eigenspace of nonnegative matrix  $A$ . The first part of the next theorem is the *Nonnegative Basis Theorem*, due to Rothblum in [46], and the second is the *Preferred Basis Theorem* due to Richman and Schneider [44]. For an  $n$ -vector  $x$ ,  $\text{supp}(x)$  denotes the *support* of the vector  $x$ , that is,  $\text{supp}(x) = \{1 \leq i \leq n \mid x_i \neq 0\}$ .

**Theorem 25** *Let  $A$  be a square nonnegative matrix with basic classes  $V_{i_1}, \dots, V_{i_m}$ . Then*

- a. *There exists a set of vectors  $\{x^1, \dots, x^m\}$  in  $E(A)$  such that*

$$\begin{aligned} x^j &\geq 0 \quad \text{and} \\ \text{supp}(x^j) &= \{i \mid i \succ V_{i_j}\}, \quad 1 \leq j \leq m. \end{aligned} \quad (2)$$

*Moreover, any such set of nonnegative vectors forms a basis of  $E(A)$ .*

- b. *There exists a basis for  $E(A)$  that satisfies in addition to (2) also*

$$(A - \rho(A)I)x^j = \sum_{V_{i_l} \succ V_{i_j}} c_{jl}x^l, \quad 1 \leq j \leq m, \quad (3)$$

*where the  $c_{jl}$  are positive.*

A basis of  $E(A)$  which satisfies (2) and (3) is called a *preferred-basis*. One preferred-basis for the matrix  $A$  in Example 21 is  $\{x^1, x^2\}$ , where

$$\begin{aligned} (x^1)^T &= [0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1] \quad \text{and} \\ (x^2)^T &= [1/20 \ 3/20 \ 1 \ 0 \ 14/15 \ 2/5 \ 2/3]. \end{aligned}$$

The existence of a nonnegative Jordan basis was also studied. Recall that a *Jordan chain* of length  $r$  corresponding to the eigenvalue  $\lambda$  of  $A$  is a sequence of  $r$  nonzero vectors  $x^0 = x, x^1 = (A - \lambda I)x, \dots, x^{r-1} = (A - \lambda I)^{r-1}x$  such that  $(A - \lambda I)^r x = 0$ . A *Jordan basis* for the algebraic eigenspace  $E(A)$  of a nonnegative matrix  $A$  is a basis of  $E(A)$  consisting of several Jordan chains corresponding to the Perron eigenvalue  $\rho(A)$ . As is well-known, there always exists a Jordan basis for  $E(A)$ . However, there does not always exist a nonnegative Jordan basis (i.e., a Jordan basis all of whose vectors are nonnegative) for the algebraic eigenspace of a nonnegative matrix  $A$ . Necessary and sufficient conditions for the existence of a nonnegative Jordan basis were given by Richman and Schneider [44]. To state the theorem, we need some definitions and notations. Let  $l_i$  denote the number of basic classes of level  $i$ . By the Rothblum Index Theorem, there is a basic class of level  $v$ , and no basic class of higher level. We denote by  $l(A)$  the sequence of levels  $(l_1, l_2, \dots, l_v)$ . Let  $j(A)$  be the non-increasing sequence  $(j_1, j_2, \dots, j_r)$  of the sizes of the Jordan blocks corresponding to  $\rho(A)$  (of course,  $j_1 = v$ ). We need also the notion of a *dual sequence*. If  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_r)$  is a sequence of non-increasing positive integers, the dual sequence  $\alpha^*$  is obtained so: Create a diagram of  $r$  left-aligned rows of dots,  $\alpha_i$  dots in the  $i$ th row from the top. Let  $\alpha^*$  be the (non-increasing) sequence of column-lengths of the diagram. For example, if  $\alpha = (4, 3, 2, 2, 1)$ , then  $\alpha^* = (5, 4, 2, 1)$ , as seen in the following diagram:



With these definitions, Richman and Schneider's theorem may now be stated.

**Theorem 26** *Let  $A$  be an  $n \times n$  nonnegative matrix. Then the following are equivalent:*

- There exists a nonnegative Jordan basis for the algebraic eigenspace  $E(A)$ .*
- For every  $1 \leq l \leq v$  there exists a nonnegative basis for  $\mathcal{N}((A - \rho(A)I)^l)$ .*
- $l(A) = j(A)^*$ .*

For the matrix  $A$  of Example 21, there are two basic classes, one of level 2 and one of level 1. Hence  $l(A) = (1, 1)$ . The index of  $A$  is 2, and the algebraic multiplicity of  $\rho(A) = 3$  is also 2, so  $j(A) = (2)$  (see also Corollary 23. Clearly



$j(A)^* = l(A)$ , so there exists a nonnegative Jordan basis for the algebraic eigenspace  $E(A)$ . Indeed, the preferred basis of  $A$  presented above is also a Jordan basis for  $E(A)$ .

A reducible nonnegative matrix may have nonnegative eigenvectors associated with eigenvalues other than the Perron eigenvalue. An eigenvalue of  $A$  is a *distinguished eigenvalue* if it has a corresponding nonnegative eigenvector. A distinguished eigenvalue is necessarily nonnegative. A class  $V_i$  in  $R(A)$  is a *distinguished class* if  $\rho(A_{ii}) > \rho(A_{jj})$  for every  $j \neq i$  such that  $V_j$  has access to  $V_i$ . In Example 21 the basic class  $V_5$  is distinguished, as is the (non-basic) class  $V_4$ , while the basic class  $V_2$  and the classes  $V_1$  and  $V_3$  are not distinguished.

The following theorem was proved by Victory [54], and was already implicit in the original paper by Frobenius [23].

**Theorem 27 (Frobenius–Victory Theorem)** *Let  $A$  be an  $n \times n$  nonnegative matrix. Then*

- A real number  $\lambda$  is a distinguished eigenvalue of  $A$  if and only if there exists a distinguished class  $V_i$  of  $A$  such that  $\rho(A_{ii}) = \lambda$ .*
- If  $V_i$  is a distinguished class of  $A$  and  $\lambda = \rho(A_{ii})$ , then there is a nonnegative eigenvector  $x$  corresponding to  $\lambda$  such that*

$$\text{supp}(x) = \{1 \leq j \leq n \mid j \geq V_i \text{ in } \Gamma(A)\}.$$

*The vector  $x$  with these properties is unique up to a positive scalar multiple.*

For the matrix  $A$  in Example 21, in addition to  $\rho(A) = 3$  the eigenvalue 0, which is  $\rho(A_{44})$ , is a distinguished eigenvalue. The only vertex with access in  $\Gamma(A)$  to  $V_4 = \{5\}$  is 5 itself, and

$$x = [0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0]^T$$

is the unique nonnegative eigenvector corresponding to 0 whose support is  $\{5\}$ .

Some remarks about the terminology: In Rothblum's paper *level* was termed *height*. Later *level* was used (see [30]), reserving the term *height characteristic* to represent the Weyr's characteristic, that is, the sequence  $\eta(A) = (\eta_1, \dots, \eta_\nu)$  with  $\eta_i = \dim \mathcal{N}((\rho(A)I - A)^i) - \dim \mathcal{N}((\rho(A)I - A)^{i-1})$ . The sequence  $l(A)$  is called the *level characteristic* of  $A$ , and the sequence  $j(A)$  of Jordan blocks is known as the *Segré characteristic*, e. g. [30]. *Non-negative bases* are referred to as *semipositive bases* in many of the works mentioned in this section, where a *semipositive* matrix or vector is a nonnegative matrix or vector which is not equal to zero.

The results here are over 20 years old – see [49], where additional references may be found. Much work was done on this subject in the years since then, which we don't cover here. In a series of papers by Tam, Tam and Wu, Tam and Schneider, the above results, and more, were extended to cone-preserving maps, providing also alternative proofs for the case of nonnegative matrices. The latest installment in this series of papers is [51], where reference can be found to the previous papers in the series, as well as to works of others. The existence of nonnegative bases, relations between characteristics of the reduced graph  $R(A)$  and characteristics of the matrix  $A$ , and extensions of such results to general (not necessarily nonnegative) matrices, were further studied by Hershkowitz, Schneider and others, see [30] and the references therein. See also Sect. 9.3 in [31], and the references there.

## Examples: Ranking Problems

### Google PageRank Revisited

We now consider in more detail the Google PageRank Example.

In the Google model, we may think of a weighted digraph  $W$  representing the links on the Web. Each of the  $N$  pages is represented by a vertex, each link is represented by an arc, and a weight is assigned to each arc: If the out-degree of a vertex  $i$  is  $d_i$ , the weight of each of the  $d_i$  arcs is  $1/d_i$ . Let  $T$  be the matrix associated with this weighted digraph, that is,  $t_{ij}$  is  $1/d_i$  if  $(i, j)$  is an arc in  $W$ , and 0 otherwise; Make  $T$  row stochastic by replacing each zero row by  $\frac{1}{N} e^T$ , where  $e$  is the  $N$ -vector with all entries equal to 1. Now  $Te = e$ . The Google matrix is:

$$P = \alpha T + (1 - \alpha) \frac{1}{N} J,$$

where  $J$  is the (in this case,  $N \times N$ ) matrix of all ones and  $\alpha = 0.85$ . By this representation of  $P$  it is easy to see that  $P$  is positive and that  $Pe = e$  (since both  $Te = \frac{1}{N}Je = e$ ). This implies (by Theorem 4 d) that 1 is the Perron root of  $P$ . Since a matrix and its transpose have the same spectrum, 1 is also the Perron root of  $P^T$ . The existence of a unique Perron vector  $\pi$  for  $P^T$  is therefore guaranteed by the Perron Theorem.

It is not known how exactly Google calculates (an approximation to)  $\pi$ , but probably some version of the power method is used. This is done essentially by starting with a vector  $x_0$  and computing  $\lim_{m \rightarrow \infty} x_m$ , where  $x_m = P^T x_{m-1} = (P^T)^m x_0$ . By Perron's Theorem (Theorem 1 e), for every vector  $x_0$  whose entries sum up to 1,  $(P^T)^m x_0$  converges to  $\pi e^T x_0 = \pi$ . Part f of the Perron Theorem gives an idea about the rate of convergence: By

a known bound for a row stochastic matrix (Chap. 5, Theorem 5.10 in [7])

$$\mu(P) \leq 1 - \sum_{j=1}^N \min_i p_{ij}.$$

The construction of  $P$  and the Web structure imply that the quantity on the right hand side is  $\alpha$ . Hence the choice of  $\alpha$  determines the rate of convergence of the power method: it is roughly the rate of convergence of  $\alpha^m$  to zero (see Theorem 1 f).

In the definition of  $P$ , the matrix  $T$  is the crucial part to the idea of Google, of ranking pages according to the hyperlink structure of the Web. It is  $T$  that contains the information on that hyperlink structure. But the matrix  $T$  is not irreducible. (The digraph representing the Web is certainly not strongly connected. Studies of the Web suggest that it has a large strongly connected component, “the core”, but also three other, equally large, groups of vertices: one consisting of vertices which have access to the core, but no access from the core; another consisting of vertices which can be accessed from the core, but have no access to the core; and finally, there is a group of vertices which have no access to or from the core [15]). The addition of the matrix  $\frac{1}{N}J$  may be explained as we did in Sect. “Matrices, Graphs and Digraphs”, but technically it has the role of making the matrix  $P$  irreducible (in fact, positive) thus guaranteeing the existence of a unique PageRank vector. If we replace the matrix  $\frac{1}{N}J$  by a matrix  $E = ev^T$ , where  $v$  is a positive probability vector (i. e.,  $v^T e = 1$ ), the new matrix  $P$  would be again positive and row stochastic. Of course, a choice of  $v \neq \frac{1}{N}e$  ( $E \neq \frac{1}{N}J$ ) means that certain pages are favored. Indeed, it is reported that Google now uses such biased perturbation matrices in the construction of  $P$ . The choice of  $\alpha$  is the result of a balancing act: large  $\alpha$  means that more weight is given to the hyperlink structure of the Web; small  $\alpha$  means faster convergence to the PageRank vector.

PageRank is not the only Web information retrieval method that is based on the hyperlink structure of the Web and uses nonnegative matrices associated with the Web digraph. See [38] for an illuminating survey on three such methods, PageRank included.

## Tournaments

Suppose that  $n$  players participate in a round-robin tournament. In such a tournament, every player plays against every other player once, and there is a clear winner in each of the games. The outcomes of the games can be represented by a digraph on  $n$  vertices, each representing a player, with  $(i, j)$  an arc if and only if player  $i$  beats

player  $j$ . Such a digraph has no loops, and for every  $i \neq j$  exactly one of  $(i, j)$  and  $(j, i)$  is an arc. It is called a *tournament*. The adjacency matrix of a tournament is called a *tournament matrix*. That is, a  $(0, 1)$ -matrix  $A$  is a tournament matrix if  $A$  if and only if  $A + A^T = J - I$ . Figure 3 shows a tournament on 4 vertices and next to it its adjacency matrix.

The problem is: Given the outcomes of all games in a round-robin tournament, how does one rank the players? It seems that it is not enough to count the scores of the players. As in the Google PageRank example, it seems reasonable to rate players not only by their scores, but also by the scores of the players they beat.

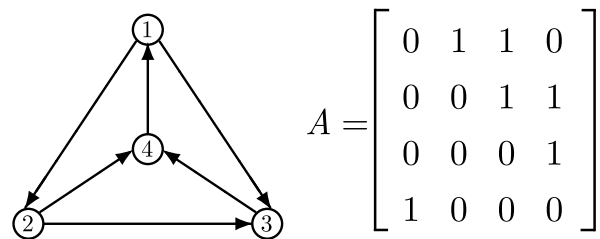
The vector of scores of the  $n$  players is  $Ae$ . The vector of second-level scores of the players, whose  $i$ th component is the sum of the scores of the players defeated by player  $i$ , is  $A(Ae) = A^2e$ . The vector of  $k$ th-level scores is  $A^k e$ . The entries of the  $k$ th-level scores grow large with  $k$ , but we are interested in their relative magnitude, so we may normalize each  $k$ th-level scores vector. By Theorem 2, if the matrix  $A$  is primitive, then

$$\lim_{k \rightarrow \infty} \left( \frac{A}{\rho(A)} \right)^k e = s,$$

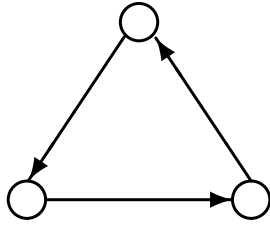
where  $s$  is a scalar multiple of the Perron vector of  $A$ . The vector  $s$  can then be used to rank the players in the tournament  $A$ : the larger  $s_i$  is, the higher is the rank of player  $i$ . For example, the tournament matrix in Fig. 4 is primitive, its Perron root is  $\rho = 1.3953$ , and  $[0.6256, 0.5516, 0.3213, 0.4484]^T$  is a corresponding eigenvector, which leads to the ranking of 1 as the top player, followed by 2, then 4, and finally 3 in the last place.

A tournament matrix need not be primitive or even irreducible; however, by the next theorem every irreducible tournament matrix of order greater than 3 is primitive.

**Theorem 28** *An  $n \times n$  irreducible tournament matrix is primitive if and only if  $n \geq 4$ .*



**Non-negative Matrices and Digraphs, Figure 4**  
A tournament and its tournament matrix



Non-negative Matrices and Digraphs, Figure 5  
A 3-cycle tournament

Tournament matrices of order 1 are 2 are reducible, and the only irreducible tournament matrices on 3 vertices are the matrices whose digraph is a 3-cycle.

All these irreducible  $3 \times 3$  tournament matrices are not primitive. It is quite obvious from Fig. 5 that in these tournaments all three players should have equal rank in any reasonable ranking method.

Here is how we may rank the players when the tournament matrix  $A$  is reducible. Let  $P$  be a permutation matrix such that

$$PAP^T = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{bmatrix}$$

is in Frobenius normal form, with each diagonal block square and irreducible (of size at least  $3 \times 3$ ), or a  $1 \times 1$  zero matrix. Let  $V_i$ ,  $i = 1, \dots, k$ , be the classes of  $A$  ( $V_i$  corresponds to  $A_{ii}$ ). Since  $A$  is a tournament matrix, each of the blocks below the diagonal has all entries equal to 1. Thus in this case  $V_i$  has access to  $V_j$  if and only if  $i > j$  (which means that each player in  $V_i$  beats each player in  $V_j$  for every  $j < i$ ). Now if  $|V_i| = 3$ , rank all the players in  $V_i$  as equal. If  $|V_i| > 3$ , rank the players in  $V_i$  according to the Perron vector of  $A_{ii}$ , and for  $i > j$  rank all players in  $V_i$  higher than the players in  $V_j$ . This is the Kendall–Wei ranking ([34,56]).

For the results of this section see Sect. 10.7 in [10]. See also [22] and the references therein for more on tournament matrices and ranking tournament players.

### The Laplacian Matrix of a Directed Graph

Let  $\Gamma$  be a digraph with no loops on vertices  $\{1, \dots, n\}$ . Let  $A = A(\Gamma)$  be the adjacency matrix of  $A$ , and let  $D = D(\Gamma)$  be the diagonal matrix of outdegrees of the vertices of  $\Gamma$  (that is,  $d_{ii}$  is the outdegree of the vertex  $i$ ). The matrix  $L(\Gamma) = D - A$  is the *Laplacian Matrix* of  $\Gamma$ . Observe that  $L(\Gamma)e = 0$ . If  $\Gamma$  is a weighted digraph, we may

replace the matrix  $A$  with the matrix  $\mathcal{A} = \mathcal{A}(\Gamma)$  whose  $i, j$  entry is the weight of the arc  $(i, j)$ , and the diagonal matrix of outdegrees by the diagonal matrix  $\mathcal{D} = \mathcal{D}(\Gamma)$  whose  $i, i$  entry is the sum of the weights of the arcs with  $i$  as initial vertex. The Laplacian of the weighted digraph  $\Gamma$  is  $\mathcal{L}(\Gamma) = \mathcal{D}(\Gamma) - \mathcal{A}(\Gamma)$ .

An *arborescence* is a digraph with no cycles, in which all the vertices but one vertex  $u$  have outdegree 1, and  $u$  has outdegree 0. The vertex  $u$  is called the *root* of the arborescence. That is, an arborescence is a digraph whose underlying graph is a tree, in which all the arcs are directed towards the root  $u$ . A *spanning arborescence* of a digraph  $\Gamma$  is a subdigraph of  $\Gamma$  with the same set of vertices as  $\Gamma$ , which is an arborescence. The Matrix-Tree Theorem below is due to Tutte [52].

**Theorem 29 (The Matrix-Tree Theorem)** *Let  $\mathcal{L}(\Gamma)$  be the Laplacian matrix of a weighted simple digraph  $\Gamma$ . Let  $\mathcal{L}_i(\Gamma)$  be the principal submatrix of  $\mathcal{L}(\Gamma)$  obtained by deleting the  $i$ th row and  $i$ th column. Then  $\det(\mathcal{L}_i(\Gamma))$  is the sum of weights of all spanning arborescences of  $\Gamma$  rooted at the vertex  $i$ , where the weight of a subdigraph is the product of the weights of its arcs.*

A non-weighted digraph  $\Gamma$ , may be considered as having weight 1 assigned to each arc. With these weights,  $\mathcal{L}(\Gamma) = L(\Gamma)$ . Thus we have:

**Corollary 30** *Let  $L(\Gamma)$  be the Laplacian matrix of a simple digraph  $\Gamma$ . Let  $L_i(\Gamma)$  be the principal submatrix of  $L(\Gamma)$  obtained by deleting the  $i$ th row and  $i$ th column. Then  $\det(L_i(\Gamma))$  is the number of spanning arborescences of  $\Gamma$  rooted at the vertex  $i$ .*

For the results here see Sect. 9.6 in [14], and Chap. VI in [53]. They are generalizations of Kirchhoff's Theorem on the (combinatorial) Laplacian of a graph. A more general form of the Matrix-Tree Theorem is the All Minors Matrix-Tree Theorem, which relates minors of order  $n - k$  of the Laplacians to sums of weights of certain spanning (directed) forests consisting of  $k$  disjoint arborescences. See e. g. [16] and the references there. See also [17] and the references there for more on such spanning forests and their relations to the Laplacian.

As mentioned in Sect. “[Matrices, Graphs and Digraphs](#)” of this article, for graphs there are many results on the connections between the spectrum of the Laplacian and the structure of the graph (see [18,21] and the references therein). There aren't many such results concerning the spectrum of the Laplacian of a digraph. Attempts in this direction have only begun in recent years, and we will touch upon these briefly in the next section, on research directions.

## Research Directions

We conclude by mentioning a few directions in current research.

Relating analytic properties of a nonnegative matrix to properties of its digraph: The Romanovsky Theorem is an (old) example of such a theorem. More recent results in this spirit can be found in [35,36,37]. In the first of these, a lower bound on the second largest eigenvalue of a stochastic matrix  $A$  is given in terms of the girth (i. e., the length of the shortest cycle) of its digraph. In the second and third, directed graphs are used in measuring the sensitivity of stationary distributions (i. e., the Perron left eigenvector of the transition matrix) under perturbations of the transition matrix.

Spectral theory for digraphs: One disadvantage of the Laplacian of a directed graph, compared to that of a graph, is that it is not symmetric and need not have real eigenvalues. Fan Chung [19] and Chai Wah Wu [57] independently derived from the Laplacian  $L = D - A$  of a digraph  $\Gamma$  a symmetric matrix whose eigenvalues may be considered instead of the eigenvalues of the original Laplacian. Following Chung, for a strongly connected digraph  $\Gamma$  let

$$S = \Phi - \frac{1}{2} (\Phi P + P^T \Phi),$$

where  $P = D^{-1}A$  and  $\Phi$  is the diagonal matrix with the left Perron vector of  $P$  on its diagonal. By scaling  $S$  we get a normalized version:

$$S = \Phi^{-1/2} S \Phi^{-1/2}.$$

The eigenvalues of  $S$  (see [57]; Wu's matrix is actually a scalar multiple of  $S$ ) or of  $S$  (see [19,20]) may be considered the directed graph equivalents of the eigenvalues of the combinatorial or normalized Laplacian of a graph. Both Chung and Wu examine connections between these eigenvalues (mainly the second smallest one) and some digraph parameters (such as the diameter and the Cheeger constant). This study seems to be at its beginning.

Searching the Web and Google: There is a lot of activity of analyzing the Google matrix, the ideas behind Google, possible improvements and alternative approaches. The best way to find the current state of affairs of these subjects is to google related terms (try that on <http://scholar.google.com>).

## Bibliography

- Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits. CRC Mathematical and Computational Biology Series. Chapman, Boca Raton
- Bang-Jensen J, Gutin G (2001) Digraphs: Theory Algorithms and Applications. Springer, London
- Bapat RB, Raghavan TES (1997) Nonnegative Matrices and Applications. Encyclopedia of Mathematics and its Applications 64. Cambridge University Press, Cambridge
- Beineke LW, Wilson RJ (2004) Topics in Algebraic Graph Theory. Encyclopedia of Mathematics and its Applications 102. Cambridge University Press, Cambridge
- Berman A (2003) Graphs of matrices and matrices of graphs. Numer Math J Chin Univ (Engl Ser) 12(suppl):12–14
- Berman A, Neumann M, Stern RJ (1989) Nonnegative Matrices in Dynamical Systems. Wiley-Interscience, New York
- Berman A, Plemmons R (1994) Nonnegative Matrices in the Mathematical sciences. SIAM, Philadelphia
- Berman A, Shaked-Monderer N (2003) Completely Positive Matrices. World Scientific, River Edge
- Birkhoff G (1946) Tres obseraciones sobre el algebra lineal. Univ Nac Tucuman Rev Ser A 5:147–150
- Bondy JA, Murty USR (1976) Graph Theory with Applications. North-Holland, New York
- Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst 33(1–7):107–117
- Brualdi RA (1979) Matrices permutation equivalent to irreducible matrices and applications. Linear Multilinear Algebra 7:1–12
- Brualdi RA, Parter SV, Schneider H (1966) The diagonal equivalence of a nonnegative matrix to a stochastic matrix. J Math Anal Appl 16:31–50
- Brualdi RA, Ryser HJ (1991) Combinatorial Matrix Theory. Encyclopedia of Mathematics and its Applications 39. Cambridge University Press, Cambridge
- Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Comput Netw 33(1–3): 309–320
- Chaiken S (1982) A combinatorial proof of the all-minors matrix tree theorem. SIAM J Alg Dis Meth 3:319–329
- Chebotaev P, Agaev R (2002) Forest matrices around the Laplacian matrix. Linear Algebra Appl 356:253–274
- Chung FRK (1997) Spectral Graph Theory. CBMS, Regional Conference Series in Mathematics 92. AMS, Providence
- Chung F (2005) Laplacians and the Cheeger inequality for directed graphs. Ann Comb 9:1–19
- Chung F (2006) The diameter and Laplacian eigenvalues of a directed graphs. Electron J Combin 13(4):6
- Cvetković D, Doob M, Sachs H (1980) Spectra of Graphs: Theory and Applications. Academic Press, New York
- Eshenbach C, Hall F, Hemansinha R, Kirkland S, Li Z, Shader B, Stuart J, Weaver J (2000) Properties of Tournaments among Well-Matched Players. Amer Math Month 107(10):881–892
- Frobenius G (1912) Über Matrizen aus nicht negativen Elementen. Sitzungsbericht. Preussische Akademie der Wissenschaften, Berlin, pp 456–477
- Gantmacher FR (1959) The Theory of Matrices. (Translated from Russian), Chelsea, New York
- Godsil C, Royle G (2001) Algebraic Graph Theory. Graduate Texts in Mathematics 207. Springer, New York
- Gulli A, Signorini A (2005) The indexable web is more than 11.5 billion pages. In: Proc. 14th WWW (Posters), pp 902–903
- Gupta RP (1967) On basis digraphs. J Combin Theory 3: 309–311



28. Hall P (1935) On representatives of subsets. *J London Math Soc* 10:26–30
29. Hartfiel DJ (1970) A simplified form for nearly reducible and nearly decomposable matrices. *Proc Amer Math Soc* 24: 388–393
30. Hershkowitz D (1999) The combinatorial structure of generalized eigenspaces – from nonnegative matrices to general matrices. *Linear Algebra Appl* 302–303:173–191
31. Hogben L (2007) *Handbook of Linear Algebra*. Discrete Mathematics and Its Applications Series 39. CRC Press, Boca Raton
32. van der Holst H, Lovász L, Schrijver A (1999) The Colin de Verdière graph parameter. In: *Graph Theory and Computational Biology*. Bolyai Soc Math Stud 7:29–85
33. Horn RA, Johnson CR (1985) *Matrix Analysis*. Cambridge University Press, Cambridge
34. Kendall MG (1955) Further contributions to the theory of paired comparisons. *Biometrics* 11:43–62
35. Kirkland S (2004) Digraph-based conditioning for Markov chains. *Linear Algebra Appl* 385:81–93
36. Kirkland SJ (2004) A combinatorial approach to the conditioning of a single entry in the stationary distribution for a Markov chain. *Electron J Linear Algebra* 11:168–179
37. Kirkland SJ (2004/5) Girth and subdominant eigenvalues for stochastic matrices. *Electron J Linear Algebra* 12:25–41
38. Langville AN, Meyer CD (2005) A survey of eigenvector methods of Web information retrieval. *SIAM Rev* 47(1):135–161
39. Lee D, Seung H (2001) Algorithms for nonnegative matrix factorization. *Adv Neural Process* 13:556–562
40. Marcus M, Minc H (1963) Disjoint pairs of sets and incidence matrices. *Illinois J Math* 7:137–147
41. Marcus M, Minc H (1964) *A survey of matrix theory and matrix inequalities*. Allyn and Bacon, Boston
42. Minc H (1988) *Nonnegative Matrices*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, New York
43. Perron O (1907) Zur Theorie der Matrizen. *Math Ann* 64: 248–263
44. Richman D, Schneider H (1978) On the singular graph and the Weyr characteristic of an  $M$ -matrix. *Aequationes Math* 17: 208–234
45. Romanovsky V (1936) Recherche sur les chaînes de Markoff. *Acta Math* 66:147–251
46. Rothblum UG (1975) Algebraic eigenspaces of nonnegative matrices. *Linear Algebra Appl* 12:281–292
47. Schneider H (1956) The elementary divisors associated with 0 of a singular  $M$ -matrix. *Proc Edinburgh Math Soc* 10(2):108–122
48. Schneider H (1977) The concepts of irreducibility and full indecomposability of a matrix in the works of Frobenius, König and Markov. *Linear Algebra Appl* 18:139–162
49. Schneider H (1986) The influence of the marked reduced graph of a nonnegative matrix on the Jordan Form and on related properties: a survey. *Linear Algebra Appl* 84:161–189
50. Seneta E (1981) *Nonnegative Matrices and Markov Chains*. Springer Series in Statistics, 2nd edn. Springer, New York
51. Tam BS (2004) The Perron generalized eigenspace and the spectral cone of a cone-preserving map. *Linear Algebra Appl* 393:375–429
52. Tutte WT (1948) The dissection of equilateral triangles into equilateral triangles. *Proc Cambridge Philos Soc* 7:463–482
53. Tutte WT (1980) *Graph Theory*. Encyclopedia of Mathematics and its Applications 21. Cambridge University Press, Cambridge
54. Victory Jr HD (1985) On nonnegative solutions to matrix equations. *SIAM J Alg Dis Meth* 6:406–412
55. Wasserman S, Faust K (1994) *Social network analysis: Methods and applications*. Cambridge University Press, Cambridge
56. Wei TH (1952) *The Algebraic Foundations of Ranking Theory*. Ph D Thesis, Cambridge University
57. Wu CW (2005) On Rayleigh-Ritz ratios of a generalized Laplacian matrix of directed graphs. *Linear Algebra Appl* 402:207–227

---

## Nonparametric Tests for Independence

CEES DIKS

University of Amsterdam, Amsterdam, The Netherlands

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Invariant Tests](#)

[Tests Based on Divergence Measures](#)

[Tests Based on Other Measures of Dependence](#)

[Bootstrap and Permutation Tests](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Hypothesis** A hypothesis is a statement concerning the (joint) distribution underlying the observed data.

**Nonparametric test** In contrast to a parametric test, a nonparametric test does not presume a particular parametric structure concerning the data generating process.

**Serial dependence** Statistical dependence among time series observations.

**Time series** A sequence of observed values of some variable over time, such as a historical temperature record, a sequence of closing prices of a stock, etc.

### Definition of the Subject

One of the central goals of data analysis is to measure and model the statistical dependence among random variables. Not surprisingly, therefore, the question whether two or more random variables are statistically independent can be encountered in a wide range of contexts. Although this article will focus on tests for independence among time series data, its relevance is not limited to the time series context only. In fact many of the dependence measures discussed could be utilized for testing independence between



random variables in other statistical settings (e.g. cross-sectional dependence in spatial statistics).

When working with time series data that are noisy by nature, such as financial returns data, testing for serial independence is often a preliminary step carried out before modeling the data generating process or implementing a prediction algorithm for future observations. A straightforward application in finance consists of testing the random walk hypothesis by checking whether increments of, for instance, log prices or exchange rates, are independent and identically distributed [8,12,80]. Another important application consists of checking for remaining dependence structure among the residuals of an estimated time series model.

## Introduction

Throughout this article it will be assumed that  $\{X_t\}$ ,  $t \in \mathbb{Z}$ , represents a strictly stationary time series process, and tests for serial independence are to be based on an observed finite sequence  $\{X_t\}_{t=1}^n$ . Unless stated otherwise, it will be assumed that the observations  $X_t$  take values in the real line  $\mathbb{R}$ . Admittedly, this is a limitation, since there are also time series processes that do not take values in the real line. For instance, the natural space in which wind direction data take values is the space of planar angles, which are naturally represented by the interval  $[0, 2\pi]$  with the endpoints identified. However, most of the tests developed to date are designed for the real-valued case. The problem under consideration is that of testing the null hypothesis that the time series process  $\{X_t\}$  consists of independent, identically distributed (i.i.d.) random variables. In practice this is tested by looking for dependence among of  $m$  consecutive observations  $X_{t-m+1}, \dots, X_t$  for a finite value  $m \geq 2$ .

Traditionally, tests for serial independence have focused on detecting serial dependence structure in stationary time series data by estimating the autocorrelation function (acf),  $\rho_k = \text{Cov}(X_{t-k}, X_t)/\text{Var}(X_t)$ , or the normalized spectral density, which is one-to-one related to the acf by Fourier transformation:

$$h(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \rho_k e^{-i\omega k}$$

and  $\rho_k = \int_{-\pi}^{\pi} h(\omega) e^{ik\omega} d\omega$ .

Because the acf is real and symmetric ( $\rho_k = \rho_{-k}$ ) the normalized spectral density is real and symmetric also. Since the acf and the normalized spectral density are related by an invertible transformation, they carry the same informa-

tion regarding the dependence of a process. For i.i.d. processes with finite variance,  $\rho_k = 0$  for  $k \geq 1$  under the null hypothesis. The spectral density is flat (equal to 1 for all  $\omega$ ) in that case.

Tests based on the acf date back to Von Neumann [83]. Motivated by the aim to test for serial independence against the presence of trends, he introduced the ratio of the mean square first difference to the sample variance,

$$S_n := \frac{\frac{1}{n-1} \sum_{t=2}^n (X_t - X_{t-1})^2}{\frac{1}{n} \sum_{t=1}^n (X_t - \bar{X})^2},$$

which may be considered a rescaled (and shifted) estimator of  $\rho_1$ . Von Neumann studied the distributional properties of this statistic in detail under the assumption of normality. Durbin and Watson [38,39] used an analogue of Von Neumann's ratio to check for first order autocorrelation among the error terms  $\{\varepsilon_t\}_{t=1}^n$  in a linear regression model, based on observed residuals  $\{\hat{\varepsilon}_t\}_{t=1}^n$ . As for the original statistic of Von Neumann, the null distribution (which is no longer unique in this case, but depends on the parameters of the data generating process) has been studied in detail for the normal case [40,56,101]. For the class of autoregressive integrated moving average (ARIMA) processes, acf-based tests for residual autocorrelation beyond lag 1 were proposed by Box and Jenkins [13] and Ljung and Box [79], and for autocorrelation in squared residuals by McLeod and Li [82]. Beran [9] proposed adapted tests for serial independence for processes with long-range dependence.

Although the autocovariance structure of a time series process fully characterizes the dependence structure within classes of linear Gaussian random processes, tests based solely on the acf may clearly fail to be consistent against dependence that does not show up in the acf. It is not hard to construct examples of processes for which this is the case. For instance, the bilinear process

$$X_t = aX_{t-2}\varepsilon_{t-1} + \varepsilon_t, \quad (|a| < 1)$$

where  $\{\varepsilon_t\}$  is a sequence of independent standard normal random variables, clearly exhibits dependence, but has no autocorrelation structure beyond lag zero. Other examples include the ARCH(1) process [42],

$$X_t = \sqrt{h_t}\varepsilon_t, \quad h_t = c + \theta X_{t-1}^2, \quad (c > 0, 0 \leq \theta < 1)$$

and the GARCH(1,1) process [11],

$$X_t = \sqrt{h_t}\varepsilon_t, \quad h_t = c + \alpha h_{t-1} + \beta X_{t-1}^2, \\ (c > 0, \alpha, \beta > 0, \alpha + \beta < 1)$$

which have become popular for modeling financial returns data.

The desire to avoid specific assumptions about the process under the null hypothesis or under the possible alternatives motivates a nonparametric statistical approach to the problem of testing for serial independence. One possibility is to develop rank-based tests for serial independence against particular types of structure such as autoregressive moving average (ARMA) structure. Compared to the linear Gaussian paradigm, this approach explicitly drops the assumption that the marginal distribution is normal, which in a natural way leads to tests formulated in terms of ranks. This has the advantage that the tests are distribution free (the null distributions of test statistics do not depend on the actual marginal distribution). The developments around invariant tests for serial independence are reviewed briefly in Sect. “Invariant Tests”.

Another nonparametric approach consists of using nonparametric estimators of divergence measures between distributions to construct tests against unspecified alternatives. The idea is to measure the discrepancy between the joint distribution of  $(X_{t-m+1}, \dots, X_t)$  and the product of marginals with a measure of divergence between multivariate probability measures. This typically involves estimating a suitable measure of dependence, and determining the statistical significance of the observed value of the statistic for the sample at hand. In Sect. “Tests Based on Divergence Measures” several tests for serial independence based on divergence measures are reviewed. For further details on some of the earlier methods, the interested reader is referred to the overview by Tjøstheim [103].

Section “Tests Based on Other Measures of Dependence” describes tests for independence based on some other measures of serial dependence in the observed time series, such as partial sums of observations and the bispectrum.

For particular statistics of interest critical values can be obtained in different ways. The traditional way is to use critical values based on asymptotic theory, which is concerned with the large sample limiting distributions of test statistics. With the increasing computer power that became available to most researchers in the recent decades, it has become more and more popular to obtain critical values of test statistics by resampling and other computer simulation techniques. I will discuss the advantages and disadvantages of several of these numerical procedures in Sect. “Bootstrap and Permutation Tests”.

Note that pairs of delay vectors such as  $(X_{t-1}, X_t)$  and  $(X_{s-1}, X_s)'$  for  $s \neq t$  may have elements in common, and hence are not independent even under the null; a fact which has to be taken into account when critical values of test statistics are determined. Tests for independence

among  $m$  random variables,  $(Y_1, \dots, Y_m)$ , say, based on a random sample therefore typically need to be adapted for applications in a time series setting. For most asymptotic results that depend on the distribution of the test statistic under the alternative (such as consistency) additional assumptions are required on the rate of decay of the dependence in the data, known as mixing conditions [14].

### Notation

Let  $X_t^m$  be short-hand notation for the delay vector  $(X_{t-m+1}, \dots, X_t)$ ,  $m \geq 3$ . For the case  $m = 2$ ,  $X_t^m$  refers to a bivariate vector  $(X_{t-k}, X_t)$ ,  $k \geq 1$ , where the value of  $k$  will be clear from the context. Under the assumption that  $X_t$  takes values in the real line  $\mathbb{R}$  (or a subset thereof) one may state the null hypothesis in terms of the joint and marginal cumulative distribution functions (CDFs):

$$H_0 : F_m(\mathbf{x}) = F_1(x_1) \times \dots \times F_1(x_m),$$

where  $\mathbf{x} = (x_1, \dots, x_m)'$ , and  $F_m(\mathbf{x}) = P(X_1 \leq x_1, \dots, X_m \leq x_m)$  is the joint cumulative distribution function (CDF) of  $X_t^m$ , and  $F_1(x) = P(X \leq x)$  the marginal CDF of  $\{X_t\}$ . If  $X_t^m$  is a continuous random variable, one can denote its probability density function by  $f_m(\mathbf{x})$ , and the independence of the elements of  $X_t^m$  can be written as

$$H_0 : f_m(\mathbf{x}) = f_1(x_1) \times \dots \times f_1(x_m),$$

where  $f_1(x)$  is the marginal probability density function of  $\{X_t\}$ . For convenience I will drop the subscript  $m$  in  $f_m(\mathbf{x})$ , and introduce  $g(\mathbf{x}) := f_1(x_1) \times \dots \times f_1(x_m)$ , so that the null hypothesis can be rephrased simply as

$$H_0 : f(\mathbf{x}) = g(\mathbf{x}).$$

### Some Practical Considerations

Which of the tests described below should one choose for practical applications? The alternative against which the null hypothesis is to be tested is any deviation from the above factorizations for some  $m \geq 2$ . Ideally, one would like a nonparametric test to have large power against all types of dependence. However, since no uniformly most powerful test against all possible alternatives exists, among the tests proposed in the literature one typically finds that some perform better against certain alternatives and some against others, and it is often hard to identify exactly why. Although power against the alternative at hand is obviously important in applications, usually these alternatives are not known in a simple parametric form. This is precisely what motivated many of the recently developed tests for independence; they

are designed to have power against large classes of alternatives. When a practitioner has to choose among the large number of the omnibus tests available, besides power also some other properties can be taken into consideration. For instance, some tests are invariant (immune to invertible transformations of the data, see Sect. “Invariant Tests”) while others, such as those based on true divergence measures discussed in Sect. “Tests Based on Divergence Measures”, are consistent against any fixed alternative, which means that they will asymptotically (with increasing sample size) detect any given alternative with probability one.

Although invariance is a pleasant property, because it allows one to tabulate the null distribution, I would generally rank consistency against any fixed alternative as more important, since invariance can usually be achieved easily by a simple trick, such as transforming the data to ranks before applying any given independence test. At the same time I should add that if one is willing to settle for power against particular classes of alternatives only, it is sometimes possible to construct an ideal hybrid between invariance and consistency in the form of an optimal invariant test. An example is the optimal rank test of Benghabrit and Hallin [7] discussed in the next section.

A clear disadvantage of omnibus tests is that after a rejection of the null hypothesis it leaves the practitioner with the problem of having to identify the type of dependence separately. If one is confident enough to assume (or lucky enough to know) a specific parametric form for the data generating process it is arguably more efficient to rely on traditional parametric methods. However, I think that in most cases the potential efficiency gains are not worth the risk of biased test results due to a misspecification of an unknown type.

## Invariant Tests

When developing nonparametric tests for serial independence it is typically assumed that the marginal distribution of the observed time series process is unknown. Because in general the distribution of the test statistic will depend on this unknown distribution, the latter plays the role of an infinite dimensional nuisance parameter. There are various ways of dealing with this problem, such as focusing on particular classes of test statistics and appealing to asymptotic theory, or using bootstrap or permutation techniques. These methods are rather common and are used in most of the tests discussed in the subsequent sections. This section is concerned with a more direct way to deal with the nuisance parameter problem. The main idea is to focus on dependence measures that are invariant un-

der one-to-one transformations of the space in which  $X_t$  takes values (so-called static transformations  $X'_t = \phi(X_t)$ , where  $\phi$  is a strictly monotonous map from  $\mathbb{R}$  to itself). This naturally leads to the study of statistics based on ranks.

## Rank Tests

Various analogues of the correlation coefficient have been proposed based on ranks. For the pairs  $\{(X_t, Y_t)\}$ , Spearman's rank correlation [97] is the sample correlation of  $R_t$  and  $S_t$ , the ranks of  $X_t$  and  $Y_t$  among the observed  $X$ 's and the  $Y$ 's, respectively. In a univariate time series context one can easily define a serial version of this rank correlation (e.g. the sample autocorrelation function of the sequence of ranks  $\{R_t\}$  of the  $X$ 's). Kendall's tau [73] for pairs  $\{(X_t, Y_t)\}$  is another rank-based measure of dependence, quantifying the concordance of the signs of  $X_i - X_j$  and  $Y_i - Y_j$ . The serial version of tau can be defined as

$$\tau_k = \binom{n-k}{2} \sum_{i=1}^{n-k} \sum_{j=1}^{i-1} \text{sgn}(X_i - X_j) \text{sgn}(X_{i+k} - X_{j+k}).$$

The multivariate versions of these concordance orderings have been described by Joe [69]. Genest et al. [45] considered tests for serial independence, building on asymptotic results derived by Ferguson et al. [43] for a serialized version of Kendall's tau in a time series setting.

Many other rank-based tests for independence have been developed meanwhile. The earlier work in this direction is covered in the review paper by Dufour [36]. Later work includes that by Bartels [6], who developed a rank-based version of Von Neumann's statistic, Hallin et al. [54] who proposed rank-based tests for serial independence against ARMA structure, and Hallin and Mélard [55] who study the finite sample behavior and robustness against outliers of their proposed procedures. Kallenberg and Ledwina [72] developed a nonparametric test for the dependence of two variables by testing for dependence of the joint distribution of ranks in a parametric model for the rank dependence.

Optimal rank tests are tests based on ranks that have maximal power. Naturally such a test depends on the alternative against which the power is designed to be large. For instance, Benghabrit and Hallin [7] derived an optimal rank test for serial independence against superdiagonal bilinear dependence.

As a way to deal with the problem that the marginal distribution is unknown under the null hypothesis (the nuisance parameter problem) Genest, Ghoudi and Rémi-lard [48] consider rank-based versions of the BDS test

statistic (see Sect. “Correlation Integrals”), as well as several other rank-based statistics.

### Empirical Copulae

As noted by Genest and Rémillard [46], a rank test for serial independence can alternatively be considered a test based on the empirical copula. The reason is that the empirical copula determines the sequence of ranks and vice versa. I therefore briefly review the notion of a copula.

If  $X_t^m$  is a continuous random variable, its copula  $C$  is defined as

$$F(x_1, \dots, x_m) = C(F(x_1), \dots, F(x_m)),$$

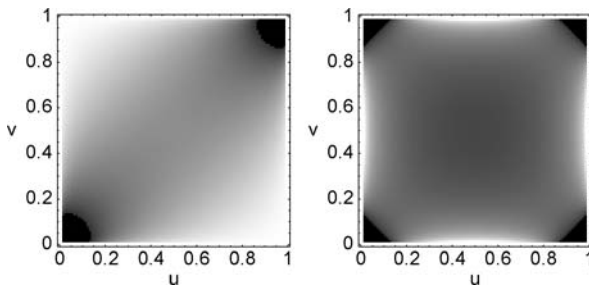
where  $F(x)$  denotes the marginal CDF. Note that  $C(u_1, \dots, u_m)$  is defined on the unit (hyper-)cube  $[0, 1]^m$  (unit square for  $m = 2$ ), and has the properties of a CDF of some distribution on that space. This allows one to define the associated copula density on the unit cube as

$$c(u_1, \dots, u_m) = \frac{\partial^m}{\partial u_1 \dots \partial u_m} C(u_1, \dots, u_m).$$

The copula density  $c$  is obtained by taking partial derivatives with respect to each of the  $X_i$ 's:

$$f(x_1, \dots, x_m) = c(F(x_1), \dots, F(x_m)) \times f(x_1) \times \dots \times f(x_m).$$

The null hypothesis of serial independence states  $f(\mathbf{x}) = g(\mathbf{x}) = f(x_1) \times \dots \times f(x_m)$ , which is equivalent to  $c(u_1, \dots, u_m) = 1$ . This shows that the factorization of the joint distribution in the product of marginals really is a property of the copula. In this sense the copula can be viewed as containing all relevant information regarding the dependence structure of  $X_t^m$ . Figure 1 shows the Gaussian copula for a bivariate distribution with correlation coefficient 0.5 and the local ARCH(1) copula (essentially a rescaled version of the copula obtained for an infinitesimal positive ARCH parameter).



**Nonparametric Tests for Independence, Figure 1**  
The Gaussian copula density for  $\rho = 0.5$  (left) and local ARCH(1) copula density (right)

The empirical copula obtained from time series data is the empirical distribution of  $(\hat{U}_{t-m+1}, \dots, \hat{U}_t)$  where  $\hat{U}_t$  is the normalized rank of  $X_t$ , defined as  $\hat{U}_t = \#\{X_s \leq X_t\}/n$ . Assuming that ties (identical values of  $X_t$  and  $X_s$  for  $t \neq s$ ) are absent, each rank only occurs once, and hence the empirical copula is one-to-one connected to the sequence of ranks. This shows that test based on ranks can be considered as tests based on the empirical copula and vice versa. It also shows that the concept of an optimal rank test against a particular copula alternative is meaningful.

The connection between sequences of ranks and empirical copulae makes it rather intuitive to design tests that have high power against serial dependence described by particular (families of) copulae. Genest and Verret [47] consider rank-based tests for independence of two random variables that are locally most powerful against a number of parametric copulae. Scaillet [92] used the copula representation to test for serial independence against positive quadrant dependence between two random variables, which holds if  $P[X \leq x, Y \leq y] \geq P[X \leq x]P[Y \leq y]$ , or equivalently  $P[X > x, Y > y] \geq P[X > x]P[Y > y]$ . In a similar vein Panchenko and I [33] derived a rank test for serial independence against the local ARCH(1) copula.

### Tests Based on Divergence Measures

In this section I consider tests for serial independence based on various dependence measures. Typically the tests obtained with this approach are not invariant. However, critical values of test statistics can still be obtained using asymptotic theory or bootstrap methods (see Sect. “Bootstrap and Permutation Tests” for more details), and the tests are consistent against a wide range of alternatives.

Many popular dependence measures are based on divergences between the  $m$ -dimensional density  $f$  and its counterpart under the null hypothesis,  $g$ . Divergences are functionals of pairs of densities, which, like distances, are equal to zero whenever  $f(\mathbf{x}) = g(\mathbf{x})$  and strictly positive otherwise. To qualify as a true distance notion between distributions a divergence measure must also be symmetric and satisfy the triangle inequality. Not all divergence measures discussed below are true distances in this sense. This is no problem for testing purposes, but if one is interested in comparing distances with other distances (e. g. for cluster analysis) then the triangle inequality is essential [81]. In general, a divergence measure might serve just as well as a distance as a basis for constructing a test for serial independence.

Tests for serial independence can roughly be divided into two groups: tests against specific types of dependence and omnibus tests, with power against general types of

dependence. For instance, the test of Von Neumann [83] mentioned above is sensitive to linear correlation between  $X_{t-1}$  and  $X_t$ , but is completely insensitive to some other types of dependence between these two variables. One of the great advantages of tests based on divergence measures is their omnibus nature. Typically these tests are consistent against any type of dependence among the  $m$  components of  $\mathbf{X}_t^m$ . Unfortunately, however, as noted in the Introduction, no uniformly most powerful test against serial independence exists, so different tests will be more powerful against different alternatives. Therefore, which test performs best in practice depends on the type of dependence structure present in the data.

Below I describe tests based on empirical distribution functions (empirical CDFs) as well as on densities. One can roughly state that tests based on the empirical CDFs are better suited to detecting large-scale deviations from the null distribution than small-scale deviations. In order for these tests to pick up deviations from independence there must be relatively large regions in  $\mathbb{R}^m$  where the density of  $\mathbf{X}_t^m$  is lower or higher than the hypothetical product density; the cumulative nature of the test statistics is relatively insensitive to small-scale deviations from the null distribution. If one wants to be able to detect subtle small-scale deviations between densities locally in the sample space, it seems more natural to use a test divergence measures based on density ratios or density differences, such as information theoretic divergences or correlation integrals. Note, however, that even among those tests performance may be hard to predict beforehand. For instance, in Subsect. “[Information Theoretic Divergence Measures](#)” I consider a family of closely related information theoretical divergence measures, but even within this family the relative powers of the tests depend strongly on the alternative at hand.

### Empirical Distribution Functions

Empirical distribution functions have been used for studying the independence of random variables at least since Hoeffding [62], who proposed dependence measures based on the difference between the joint distribution function and the product of marginals,  $F(\mathbf{x}) - G(\mathbf{x})$ , where  $G(\mathbf{x}) = \prod_{i=1}^m F_1(x_i)$  is the joint CDF under the null hypothesis.

There are various ways to define divergences in terms of distribution functions. A popular class of statistics is obtained by considering divergence measures of the type

$$d_w^2 = \int_{\mathbb{R}^m} (F(\mathbf{x}) - G(\mathbf{x}))^2 w(F(\mathbf{x})) dF(\mathbf{x}),$$

where  $w(\cdot)$  is a positive weight function. For  $w = 1$  this divergence is known as the Cramér–von Mises criterion, which has become a well-known criterion in univariate one- and two-sample problems. The Cramér–von Mises criterion suggests testing the independence of the elements of  $\mathbf{X}_t^m$  based on its sample version

$$\tilde{n}d_n^2 = \sum_{i=m}^n \left( \hat{F}(\mathbf{X}_i^m) - \hat{G}(\mathbf{X}_i^m) \right)^2,$$

where  $\tilde{n} = n - m + 1$  is the number of  $m$ -dimensional delay vectors,  $\hat{F}$  is the empirical joint CDF and  $\hat{G}(\mathbf{x}) = \prod_{i=1}^m \hat{F}_1(x_i)$  is the product of marginal empirical CDFs. This statistic, referred to as the Cramér–von Mises statistic, was proposed by Hoeffding [62] for testing independence in the bivariate case, based on a random sample of variables  $(X_i, Y_i)$  from a bivariate distribution (i.e. outside the time series scope). Since the statistic is invariant under one-to-one transformations of marginals, the tests based on it are automatically distribution-free. Although the null distribution for a random sample was known to converge in distribution to a mixture of scaled  $\chi^2(1)$  random variables,

$$\tilde{n}d_n^2 \xrightarrow{d} \frac{1}{\pi^4} \sum_{j,k=1}^{\infty} \frac{1}{j^2 k^2} Z_{jk}^2,$$

where the  $Z_{jk}$  are independent standard normal random variables, it was initially not available in a form suitable to practical applications. The distribution was tabulated eventually by Blum et al. [10], who also considered higher-variate versions of the test.

By generalizing results of Carlstein [23], Skaug and Tjøstheim [95] extended the test of Hoeffding, Blum, Kiefer and Rosenblatt to a time series context and derived the null distribution of the test statistic for first-order dependence ( $m = 2$ ) under continuous as well as discrete marginals. In the continuous case it turns out that the first order test statistic has the same limiting null distribution as for a random sample from a bivariate distribution with independent marginals. Skaug and Tjøstheim [95] also showed that the statistic  $nG_{K,n} = n \sum_{k=1}^K d_{k,n}^2$ , where  $d_{k,n}^2$  is the Cramér–von Mises statistic for  $(X_{t-k}, X_t)$ , has a mixture of scaled  $\chi^2(K)$  distributions as its limiting distribution. For high lags and moderate sample sizes Skaug and Tjøstheim [95] report that the asymptotic approximation to the finite sample null distribution is poor, and suggest the use of bootstrap methods to obtain critical values.

Delgado [29] considered the analogue test for higher-variate dependence in time series. In that case differences with the Hoeffding–Blum–Kiefer–Rosenblatt asymptotics



arise due to the presence of dependence across the delay vectors constructed from the time series. Delgado and Mora [28] investigated the test for first order independence when applied to regression residuals, and found that the test statistic in that case has the same limiting null distribution as for serially independent data.

The Kolmogorov–Smirnov statistic

$$\sqrt{n} \sup_x |\hat{F}(\mathbf{x}) - \hat{G}(\mathbf{x})|.$$

is another popular test statistic for comparing empirical cumulative distribution functions. Ghoudi et al. [49] developed asymptotic theory for this test statistic in rather general settings, which include the time series context. In their power simulations against several alternatives, including AR(1) and nonlinear moving average processes, the Cramér–von Mises statistic displayed a better overall performance than the Kolmogorov–Smirnov statistic.

This suggests that the Cramér–von Mises statistic might be a good choice for practical applications, provided that one wishes to compare empirical distribution functions. As noted above, for detecting subtle density variations it might be more suitable to use a dependence measure based on integrated functions of densities, described in the next subsections.

### Integrated Functions of Density Differences

For the bivariate case, Rosenblatt [89] and Rosenblatt and Wahlen [90] considered a class of measures of dependence based on integrated squared differences of densities

$$d(f, g) = \int_{\mathbb{R}^2} w(\mathbf{x})(f(\mathbf{x}) - g(\mathbf{x}))^2 d\mathbf{x},$$

for some positive weight function  $w(\mathbf{x})$ . The integral can be estimated nonparametrically by plugging in kernel density estimators for the unknown densities, and performing the integration, either numerically or, if possible, analytically. Alternatively one may estimate the integral by taking sample averages of estimated densities. For instance, if  $w = 1$  one may estimate  $\int_{\mathbb{R}^2} f^2(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^2} f(\mathbf{x}) dF(\mathbf{x}) = E[f(\mathbf{X}_t)]$  as  $\tilde{n}^{-1} \sum_t \hat{f}(\mathbf{X}_t)$ , where  $\hat{f}(\mathbf{x})$  represents a consistent kernel density estimate of  $f(\mathbf{x})$ .

Chan and Tran [24] proposed a test for serial independence based on the integrated absolute difference

$$\tilde{d}(f, g) = \int_{\mathbb{R}^2} |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x}, \quad p > 0,$$

for which they developed a histogram estimator [94]. They obtained critical values of the test statistic using a bootstrap method.

Skaug and Tjøstheim [96] explored tests for serial independence based on several dependence measures which are weighted integrals of  $f(\mathbf{x}) - g(\mathbf{x})$  in the bivariate case ( $m = 2$ ) including the above two measures, which they refer to as  $I_3$  and  $I_2$ , respectively. In addition they consider the Kullback–Leibler divergence ( $I_1$  in their notation, discussed in Subsect. “Information Theoretic Divergence Measures”) and

$$I_4 = \int_{\mathbb{R}^2} (f(\mathbf{x}) - g(\mathbf{x}))f(\mathbf{x}) d\mathbf{x}.$$

The latter measure is not a true divergence between  $f$  and  $g$ , but if  $f$  is a bivariate normal pdf,  $I_4 \geq 0$  with equality if and only if  $f = g$ . Skaug and Tjøstheim [96] performed a simulation study in which the corresponding estimators  $\hat{I}_i$  were compared, and  $\hat{I}_4$  was found to perform well relative to the other statistics. They subsequently investigated some of the asymptotic properties of this estimator, establishing, among other results, its asymptotic normality. Despite these encouraging simulation results one should beware that there are theoretical cases with dependence where  $I_4$  is zero, meaning that there are also processes with dependence against which the test has little or no power.

### Information Theoretic Divergence Measures

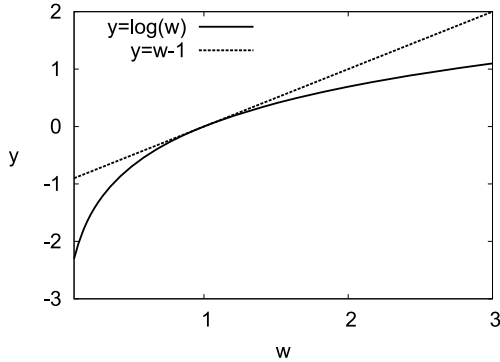
By using test statistics based on true divergences, tests can be obtained that are consistent against all deviations of  $f$  from the product measure  $g$ . Although this does not guarantee high finite sample power for specific alternatives, it obviously is a desirable property if the nature of the alternative is unknown.

Joe [68] described several information theoretic divergence measures, including the Kullback–Leibler divergence between two densities  $f$  and  $g$ , defined as

$$I(f, g) = \int_{\mathbb{R}^m} f(\mathbf{x}) \log \left( \frac{f(\mathbf{x})}{g(\mathbf{x})} \right) d\mathbf{x}.$$

In the case where  $f$  is a bivariate density, of  $X_{t-k}$  and  $X_t$ , say, and  $g$  is the product of marginal densities,  $I(f, g)$  is also known as the mutual information between  $X_{t-k}$  and  $X_t$ .

Robinson [88] took the Kullback–Leibler divergence as a starting point for testing the equivalence of  $f$  and  $g$ . The Kullback–Leibler divergence is invariant under transformations of marginal distributions, and satisfies  $I(f, g) \geq 0$  with equality if and only if  $f = g$ . To see why, consider the random variable  $W = g(\mathbf{X})/f(\mathbf{X})$ . By construction  $E[W] = 1$ , and because  $\log(W)$  is a concave function of  $W$  it follows from Jensen’s inequality that  $E[-\log W] \leq 0$ ,



**Nonparametric Tests for Independence, Figure 2**

**Illustration of Jensen's inequality.** Since the function  $y = \log w$  is concave, it is bounded from above by the tangent line at  $w = 1$ , given by  $y = w - 1$ . It follows that if  $E[W] = 1$  and  $Y = \log W$ ,  $E[Y] = E[\log W] \leq E[W - 1] = E[W] - 1 = 0$  with equality if and only if  $W = 1$  with probability 1

with equality if and only if  $g(X) = f(X)$  with probability one. The reason is that  $\log x \leq 1 - x$  for positive  $x$ , as illustrated in Fig 2. Application of this inequality to  $W$  shows that  $E[\log(W)] \leq \log E[W] = 0$  with equality if and only if  $W = 1$  with probability 1.

The fact that the Kullback–Leibler divergence is positive for any difference between the true pdf  $f$  and the hypothetical pdf  $g$  makes it a suitable quantity for testing  $f = g$  against unspecified alternatives. A consistent estimator for  $I(f, g)$  may serve to construct a test that is consistent (i. e. asymptotically rejects with probability one) against any fixed alternative. Robinson [88] proceeded by constructing such an estimator for  $I(f, g)$  using plug-in density estimates of the unknown bivariate densities  $f(\mathbf{x})$  and  $g(\mathbf{x})$ . For instance, one may use the Nadaraya–Watson density estimator

$$\hat{f}(\mathbf{x}) = \frac{1}{\tilde{n}h^m} \sum_t K((\mathbf{x} - \mathbf{X}_t)/h),$$

where  $K(\mathbf{x})$  is a probability kernel, such as the pdf of a multivariate normal random variable with independent elements,  $(2\pi)^{-m/2} \exp(-\mathbf{x}'\mathbf{x}/2)$ ,  $h$  is a smoothing parameter, and  $\tilde{n}$  the number of delay vectors  $\mathbf{X}_t$  occurring in the summation. The resulting estimator of the Kullback–Leibler divergence is

$$\hat{I}(\hat{f}, \hat{g}) = \frac{1}{|S|} \sum_{t \in S} \log \left( \frac{\hat{f}_{X_{t-k}, X_t}(X_{t-k}, X_t)}{\hat{f}_X(X_{t-k}) \hat{f}_X(X_t)} \right), \quad (1)$$

where  $S$  is a subset of  $k + 1, \dots, n$ , introduced to allow for “trimming out” some terms of the summation if desired, for instance terms in the summation for which one

or more of the local density estimates are negative or zero, as may happen depending on the type of density estimators used. The number of elements of  $S$  is denoted by  $|S|$ . Robinson [88] showed that although the test statistic is a consistent estimator of the Kullback–Leibler divergence, no scaled version of it has a standard normal limit distribution, preventing the development of asymptotic distribution theory in a standard fashion. To overcome this problem, instead of deriving the asymptotic distribution of  $\hat{I}(\hat{f}, \hat{g})$ , Robinson showed asymptotic normality of a modified test statistic, obtained by attaching weights to each of the terms in the sum in (1).

Hong and White [65] argued that this modification leads to the loss of asymptotic local (i. e. close to the null) power, and developed asymptotic distribution theory for the estimator  $\hat{I}(\hat{f}, \hat{g})$  directly. After adjusting for the asymptotic mean they found that an appropriately scaled version of the test statistic actually does have an asymptotically standard normal distribution under the null hypothesis.

Alternatively one may obtain critical values by calculating the test statistic for a large number of simulated replications of an i.i.d. process, as done by Granger and Lin [50]. Note, however, that the critical values thus obtained will depend on the marginal distribution assumed for the process. This approach was followed by Dionísio et al. [35], who used the mutual information between  $X_{t-k}$  and  $X_t$  for a range of  $k$ -values to test for serial independence in stock index returns. Critical values of the test statistic were determined by constructing a reference distribution of the test statistic under the null hypothesis by simulation, repeatedly calculating the value of the test statistic for a large number of independently generated i.i.d. normal time series. The results suggest the presence of residual dependence at several lags for log-returns on stock indices that were filtered to account for ARMA and GARCH structure.

A closely related information theoretic approach has been described by Granger et al. [51] and Racine and Maaoui [86], who start by considering the class of divergences based on the asymmetric  $q$ -class entropy divergence measure defined as

$$I_q(f, g) = \frac{1}{1-q} \left[ 1 - \int_{\mathbb{R}^m} \left( \frac{f(\mathbf{x})}{g(\mathbf{x})} \right)^q g(\mathbf{x}) d\mathbf{x} \right].$$

This is a generalization of the Kullback–Leibler divergence, which is recovered in the limit as  $q \rightarrow 1$ . The au-

thors subsequently focused on the symmetric  $q = \frac{1}{2}$  case,

$$\begin{aligned} I_{\frac{1}{2}}(f, g) &= 2 - 2 \int_{\mathbb{R}^m} \sqrt{g(\mathbf{x})} \sqrt{f(\mathbf{x})} d\mathbf{x} \\ &= \int_{\mathbb{R}^m} \left( \sqrt{g(\mathbf{x})} - \sqrt{f(\mathbf{x})} \right)^2 d\mathbf{x}, \end{aligned}$$

known as the Hellinger distance, and used this to develop tests for various hypotheses involving the equality of two densities, including serial independence.

Fernandes and Neri [44] proposed using an estimator of the Tsallis entropy [105] to test for serial independence in a time series setting. As it turns out, the Tsallis entropy is identical to  $I_q(f, g)$ . In numerical simulation studies Fernandes and Neri [44] found that, depending on the time series processes under consideration and on the value of  $q$ , these tests can have higher power than the entropy-based test of Hong and White [65]. In comparison with the BDS test of Brock et al. [16] (see Sect. “Correlation Integrals”) they found that the entropy-based tests perform worse in most cases, although the latter have more power for specific processes, including fractional AR(1) and threshold AR(1) processes.

Aparicio and Escribano [1] developed further tests based on information theoretic dependence measures. Their framework allows testing for short memory against long memory, as well as for the absence of cointegration against linear or nonlinear cointegration. In empirical applications they found that although the rates of the Peseta against the Yen and the US dollar do not appear to be linearly cointegrated, there is evidence supporting a nonlinear cointegrating relation between the two rates.

### Characteristic Functions

Csörgő [26] noted that instead of investigating empirical distribution functions for testing independence, as Hoffding [62] and Blum, Kiefer and Rosenblatt [10] did, a parallel approach can be based on empirical characteristic functions. Several tests for independence have been developed on the basis of this principle. I will be concerned here only with serial independence tests. As with the empirical distribution function one might consider various measures of deviations from independence, e. g. based on a maximum difference or on weighted integrals.

The test of Pinkse [84] is based on the observation that the random variables  $X_1$  and  $X_2$  are independent if and only if their joint characteristic function factorizes. He proposed to test the relation

$$\Psi(u, v) = Ee^{i(uX_1 + vX_2)} - Ee^{iuX_1} Ee^{ivX_2} = 0$$

through a quantity of the form  $\theta = \iint g(u)g(v)|\Psi(u, v)|^2 du dv$ , where  $g(\cdot)$  is a positive function. In fact Pinkse introduced an estimator of a related but different functional, as detailed in Sect. “Quadratic Forms” where it is also explained why the test statistic can be estimated directly using U-statistics [93], without the need to actually perform the transformation to characteristic functions.

Hong [63,64] proposed a test for independence based on the difference between the joint characteristic function of  $X_{t-j}$  and  $X_t$  and the product of their marginal characteristic functions. The main idea is to weigh the discrepancy between  $F$  and  $G$  across all lags by considering the Fourier transforms

$$h(\omega, \mathbf{x}) := (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \gamma_j(\mathbf{x}) \exp(-ij\omega)$$

of  $\gamma_j(\mathbf{x}) = F_j(\mathbf{x}) - G(\mathbf{x})$  where  $F_j$  denotes the joint CDF of  $X_{t-j}$  and  $X_t$ . An application [63] to a series of weekly Deutschmark US dollar exchange rates from 1976 until 1995 showed that although the log returns are serially uncorrelated, there is evidence of nonlinear dependence of the conditional mean return given past returns.

### Correlation Integrals

Correlation integrals have been used extensively in the chaos literature, where they were introduced to characterize deterministic dynamics reconstructed from time series. The interested reader is referred to Takens [99] for details of the reconstruction theorem, and to Grassberger et al. [52,53] and the book by Tong [104] for a snapshot of the early developments around correlation integrals. Correlation integrals turn out to be very suitable also in stochastic contexts. They are well adapted to testing for serial independence against unspecified alternatives, as shown below. Moreover, since they are U-statistics asymptotic theory is readily available for them [30,31].

Brock et al. [15,16] based their test for serial independence on the correlation integral of  $X_t^m$ , defined as

$$\begin{aligned} C_m(\varepsilon) &= P[|Z_1 - Z_2| \leq \varepsilon] \\ &\text{with } Z_i \sim X_t^m, \text{ independent for } i = 1, 2, \end{aligned}$$

where  $|\cdot|$  denotes the supremum norm defined by  $|\mathbf{x}| = \sup_{i=1, \dots, m} |x_i|$ . Under the null hypothesis of serial independence the correlation integral factorizes:

$$C_m(\varepsilon) = (C_1(\varepsilon))^m. \quad (2)$$

This can be seen by expressing  $C_m(\varepsilon)$  as a double integral

$$\begin{aligned} C_m(\varepsilon) &= \int_{\mathbb{R}^m} \int_{\mathbb{R}^m} I_{[0,\varepsilon]}(|\mathbf{x} - \mathbf{y}|) \mu_m(d\mathbf{x}) \mu_m(d\mathbf{y}) \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} I_{[0,\varepsilon]}(|x_1 - y_1|) \mu_1(dx_1) \mu_1(dy_1) \times \dots \\ &\quad \times \int_{\mathbb{R}} \int_{\mathbb{R}} I_{[0,\varepsilon]}(|x_m - y_m|) \mu_1(dx_m) \mu_1(dy_m) \\ &= (C_1(\varepsilon))^m. \end{aligned}$$

In the first step the independence of  $|X_i - Y_i|$  and  $|X_j - Y_j|$  ( $1 \leq i \neq j \leq m$ ) was used, for two vectors  $\mathbf{X}$  and  $\mathbf{Y}$  drawn independently from the distribution of  $\mathbf{X}_t^m$  under the null hypothesis (in that case all elements  $X_1, \dots, X_m, Y_1, \dots, Y_m$  are independent and identically distributed). Note that strictly speaking  $C_m(\varepsilon) - (C_1(\varepsilon))^m$  is not a divergence. Although it will typically be nonzero for most alternatives with serial dependence, it is possible to construct examples where  $C_m(\varepsilon) - (C_1(\varepsilon))^m$  is zero even under serial dependence. Formally this means that testing if  $C_m(\varepsilon) - (C_1(\varepsilon))^m$  is zero, amounts to testing an implication of the null hypothesis of serial independence.

For a given kernel function  $K(\mathbf{x}_1, \dots, \mathbf{x}_k)$  that is symmetric in its arguments, the U-statistic based on a (possibly dependent) sample  $\{\mathbf{X}_t^m\}_{t=1}^{\tilde{n}}$ , consists of the sample average of the kernel function with all elements different:

$$\frac{(\tilde{n} - k)!}{\tilde{n}!} \sum_{i_1} \dots \sum_{i_k} K(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_k}).$$

$i_j$  all different

The corresponding V-statistic is the sample average if the elements are allowed to be identical:

$$\frac{1}{\tilde{n}^k} \sum_{i_1} \dots \sum_{i_k} K(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_k}).$$

The BDS test is based on the scaled difference between the U-statistic estimators of the left- and right-hand sides of (2):

$$W_n = \sqrt{n} \frac{C_{m,n}(\varepsilon) - (C_{1,n}(\varepsilon))^m}{\sigma_{m,n}},$$

where the U-statistic

$$C_{m,n}(\varepsilon) = \frac{2}{(n - m + 1)(n - m)} \cdot \sum_{i=2}^{n-m+1} \sum_{j=1}^i I_{[0,\varepsilon]}(|\mathbf{X}_i^m - \mathbf{X}_j^m|), \quad (3)$$

is known as the sample correlation integral at embedding dimension  $m$  and  $\sigma_{m,n}^2$  is a consistent estimator of the

asymptotic variance of the scaled difference. The asymptotic distribution of the test statistic can be derived using the results for U-statistics for weakly dependent processes, described by Denker and Keller [30,31]. Under the null hypothesis of serial independence,

$$W_n \xrightarrow{d} N(0, 1).$$

In fact the asymptotic distribution of  $C_{m,n}(\varepsilon) - (C_{1,n}(\varepsilon))^m$  is obtained from that of  $(C_{m,n}(\varepsilon), C_{1,n}(\varepsilon))$ . Since this is a pair of U-statistics, it follows from the results of Denker and Keller [30] that it is asymptotically bivariate normally distributed for strongly mixing stationary processes [14]. After deriving the asymptotic means and covariance matrix one can apply the functional delta method to obtain the asymptotic normal distribution of  $C_{m,n}(\varepsilon) - (C_{1,n}(\varepsilon))^m$ .

To apply the BDS test the user should specify a value for the bandwidth parameter  $\varepsilon$ . In numerical studies as well as applied studies,  $\varepsilon$ -values are typically taken in the range 0.5–1.5 times the sample standard deviation of the observed time series. Note that the null hypothesis tested is independence among all elements of  $\mathbf{X}_t^m$  rather than pairwise independence of  $X_{t-m+1}$  and  $X_t$ . Because this results in a relative error of the estimated correlation integral that increases rapidly with  $m$ , for applications with moderate sample sizes ( $n \approx 1000$ , say) small values of  $m$  are recommendable (e.g.  $m = 2$  or  $m = 3$ ).

Brock et al. [16] derived a ‘nuisance parameter theorem’ for the BDS test, showing that the limiting distribution of the test statistic is asymptotically free of estimation uncertainty of an unknown parameter  $\theta$  (e.g. a vector of AR( $p$ ) model parameters) provided that a root- $n$  consistent estimator is available for  $\theta$ . The nuisance parameter theorem, which covers the parameters of AR models, but not, for instance of ARCH models, states that the asymptotic distribution of the test statistic for residuals is the same as that for the true innovations. This justifies the use of residuals in place of true innovations asymptotically, which is convenient since it allows using the BDS test on residuals as a model specification test, provided that the estimated parameters are root- $n$  consistent.

De Lima [78] formulated five conditions under which the BDS test is asymptotically nuisance parameter free (i.e. can be used as a model specification test). These involve, among others, mixing conditions and conditions ensuring the consistency of parameter estimates. Interestingly, the test is not asymptotically nuisance parameter free for GARCH residuals, but it is when applied to logarithms of the squared residuals. Caporale et al. [21] have performed a simulation study to evaluate the behavior of the test

statistic under violations of these conditions, and found the BDS test to be very robust.

Note that filtering time series data by replacing them with the (standardized) residuals of a time series model typically has the effect of whitening the data, which makes the detection of dependence more difficult. Brooks and Heravi [18] found that upon filtering data through a completely misspecified GARCH model, the frequency of rejection of the i.i.d. null hypothesis can fall dramatically. Therefore, a failure to reject the null hypothesis on the basis of GARCH residuals does not imply that a GARCH model is consistent with the data.

Wolff [106] observed that the unnormalized correlation integral, i. e. the double sum in (3) without the normalizing factor, converges to a Poisson law under some moderate assumptions regarding the marginal distribution. This motivates a nonparametric test procedure based on the correlation integral, which Wolff found to have reduced size distortion compared to the usual BDS test.

Instead of the sample correlation integral, Kočenda and Briatka [74] suggest using an estimator of the slope

$$D_m(\varepsilon) = \frac{d \ln C_m(\varepsilon)}{d \ln \varepsilon},$$

also known as the course-grained correlation dimension at embedding dimension  $m$  and distance  $\varepsilon$ , for testing the null hypothesis of serial independence. The intuition is that the theoretically  $C_m(\varepsilon) \sim \varepsilon^m$  for small  $\varepsilon$  under the i.i.d. null, while  $C_m(\varepsilon) \sim \varepsilon^\alpha$  for some  $\alpha < m$ , provided  $m$  is sufficiently large, in the case of a low-dimensional attractor with correlation dimension  $\alpha$ . The coarse-grained correlation dimension is a measure for complexity, and deviations from the null other than chaos typically also reduce the coarse-grained correlation dimension. This makes the coarse-grained correlation dimension a promising quantity for testing the i.i.d. null hypothesis. Rather than using the slope for a single bandwidth  $\varepsilon$ , Kočenda and Briatka [74] proposed to use an estimator of the average slope across a range of  $\varepsilon$ -values, consisting of the least squares estimator of the slope parameter  $\beta_m$  in the regression

$$\ln(C_{m,n}(\varepsilon_i)) = \alpha_m + \beta_m \ln(\varepsilon_i) + u_i, \quad i = 1, \dots, b,$$

where  $\alpha_m$  is an intercept,  $u_i$  represents an error term, and  $b$  is the number of bandwidths  $\varepsilon_i$  taken into consideration. They then determined the optimal range of  $\varepsilon$ -values by simulation. The test based on the resulting least squares estimator  $\hat{\beta}_m$  for  $\beta_m$  was found to have high power compared to some other tests for serial independence, and to behave well when used as a specification test.

Although it is clear that the correlation integrals from more than one bandwidth value  $\varepsilon_i$  contain more information than that from a single bandwidth, it is not clear why it would be a good idea to base a test on the estimator  $\hat{\beta}_m$ . Since the correlation integral is an empirical CDF (of inter-point distances) the error terms  $u_i$  will be correlated, which typically leads to a loss of efficiency. In Subsect. “Multiple Bandwidth Permutation Tests” I discuss an alternative way to combine information from different bandwidths into a single test statistic, inspired by the rate-optimal adaptive tests of Horowitz and Spokoiny [67].

Johnson and McLelland [70,71] proposed a variation on the BDS test for testing the independence of a variable and a vector based on correlation integrals. The main idea is to test for remaining dependence between residuals and regressors, in addition to mere dependence among residuals. This might be an advisable approach in many cases, because even though theoretically a model misspecification should lead to residuals with serial dependence, it is often very hard to detect this dependence with tests on the residuals only, due to the whitening effect of the filtering.

## Quadratic Forms

Quadratic forms are convenient for defining squared distances between probability distributions, which provide tests that are consistent against any type of dependence (hence including, for instance, ARCH and GARCH structure). A comparative advantage relative to the information theoretical divergences discussed in Subsect. “Information Theoretic Divergence Measures” is that they can, like correlation integrals, be estimated straightforwardly by U- and V-statistics.

The starting point for the construction of a quadratic form is a bilinear form, which may be interpreted as an inner product on the space of measures on  $\mathbb{R}^m$ . The quadratic forms discussed here were first applied in the context of testing for symmetries of multivariate distributions [34], and later extended to a time series context [32].

Consider, for a kernel function  $K(\cdot, \cdot)$  on  $\mathbb{R}^m \times \mathbb{R}^m$  the form

$$(\mu, \nu) = \int_{\mathbb{R}^m} \int_{\mathbb{R}^m} K(\mathbf{x}, \mathbf{y}) d\mu(\mathbf{x}) d\nu(\mathbf{y})$$

for measures  $\mu$  and  $\nu$ . Note that this form is bilinear (linear in  $\mu$  as well as  $\nu$ ). If this form happens to satisfy  $(\mu, \mu) \geq 0$  for any (possibly signed) measure  $\mu$  with  $(\mu, \mu) = 0$  if and only if  $\mu(A) = 0$  for all Borel subsets  $A$  of  $\mathbb{R}^m$ , then  $K$  is called positive definite. In the terminology introduced above, this means that  $(\mu - \nu, \mu - \nu)$  is a divergence between the measures  $\mu$  and  $\nu$ . Note that a positive definite



form defines an inner product on the space of measures on  $\mathbb{R}^m$  with the usual properties:

- (i)  $(\mu, \nu) = (\nu, \mu)$ .
- (ii)  $(a\mu + b\nu, \eta) = a(\mu, \eta) + b(\nu, \eta)$  for scalars  $a, b$ .
- (iii)  $(\mu, \mu) \geq 0$  with equality iff  $\mu(A) = 0$  for any Borel subset  $A \in \mathcal{A}$ .

The inner product can therefore be used to define a norm of  $\mu - \nu$  as  $\|\mu - \nu\| = \sqrt{(\mu - \nu, \mu - \nu)}$ , which satisfies all the usual properties of a distance, such as Schwarz's inequality, the triangle inequality and the parallelogram law (See e.g. Debnath and Mikusiński [27]).

In short: any positive definite kernel  $K$  defines an inner product on the space of measures on  $\mathbb{R}^m$ , which in turn defines a squared distance between  $\mu$  and  $\nu$ , given by

$$\theta = \|\mu - \nu\|^2 = (\mu - \nu, \mu - \nu).$$

(For simplicity the dependence of the squared distance on the kernel function  $K$  has been suppressed in the notation.)

To pinpoint some classes of kernel functions that are suitable for our purposes (i. e. that are positive definite) let us assume that the kernel function  $K$  depends on  $x$  and  $y$  only through the difference  $x - y$ , and that the kernel function factorizes, i. e.

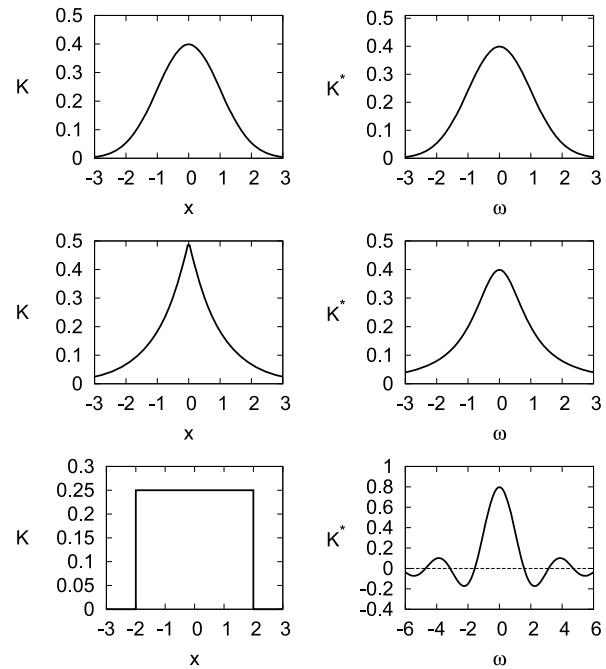
$$K(x, y) = \prod_{i=1}^m \kappa(x_i - y_i).$$

In that case the Fourier transform  $\tilde{K}$  of the kernel function also factorizes, into  $\tilde{K}(\mathbf{u}) = \prod_{i=1}^m \tilde{\kappa}(u_i)$ , where  $\tilde{\kappa}(u) = \int \kappa(t)e^{-iut}dt$ , the Fourier transform of  $\kappa$ . The squared distance  $\theta = \|\mu - \nu\|^2$  can then be expressed directly in terms of characteristic functions  $\tilde{\mu}$  and  $\tilde{\nu}$  of  $\mu$  and  $\nu$  respectively:

$$\theta = \frac{1}{2\pi} \int \tilde{K}(\mathbf{u}) |\tilde{\mu}(\mathbf{u}) - \tilde{\nu}(\mathbf{u})|^2 d\mathbf{u}.$$

This follows from applying Parseval's theorem to  $\theta = \iint K(\mathbf{x}, \mathbf{y})(\mu - \nu)(d\mathbf{x})(\mu - \nu)(d\mathbf{y})$ . It follows that if the kernel function is bounded and has a Fourier transform which does not vanish on any interval, its associated bilinear form is positive definite.

To illustrate this, Fig. 3 shows three kernel functions and their Fourier transforms. The Gaussian kernel (top panels) has a Gaussian as its Fourier transform, which is everywhere positive. Therefore, the Gaussian product kernel is positive definite and defines a quadratic form suitable for detecting any differences between a pair of distributions on  $\mathbb{R}^m$ . A similar conclusion holds for the



Nonparametric Tests for Independence, Figure 3

Kernel functions (left) and their Fourier transforms (right) for the Gaussian kernel (top), double exponential kernel (middle) and the naive kernel (bottom)

double exponential kernel  $\exp(-|x|/a)$  (middle panels). The 'naive' kernel function  $I_{[-a,a]}(x)$  (bottom panels) has a Fourier transform which is negative for certain ranges of the frequency  $\omega$ , and hence is not a positive definite kernel function.

Given the kernel function  $K$  (e.g. a multivariate Gaussian product kernel) the estimation of the associated quadratic form  $(\mu - \nu, \mu - \nu) = (\mu, \mu) - 2(\mu, \nu) + (\nu, \nu)$  is straightforward. Empirical versions of  $(\mu, \mu)$ ,  $(\mu, \nu)$  and  $(\nu, \nu)$  can be obtained easily as sample averages. For instance, if  $\mathbf{X}_i^m$  is a sample from  $\mu$ , the sample version of  $(\mu, \mu) = \iint K(\mathbf{s}_1, \mathbf{s}_2)d\mu(\mathbf{s}_1)d\mu(\mathbf{s}_2)$  is the V-statistic

$$\widehat{(\mu, \mu)} = \frac{1}{\tilde{n}^2} \sum_i \sum_j K(\mathbf{X}_i^m, \mathbf{X}_j^m).$$

As before,  $\tilde{n} = n - m + 1$  denotes the number of  $m$ -vectors available. It follows from the results of Denker and Keller [30,31] for U- and V-statistics of dependent processes that the estimator is consistent under strong mixing conditions and asymptotically normally distributed with a variance that can be estimated consistently from the data. Note that the estimator of  $(\mu, \mu)$  is in fact a sample correlation integral, but with the kernel  $K$  instead of the usual naive kernel.

As shown in [32], similar consistent estimators for the other terms can be constructed easily:

$$\begin{aligned}\widehat{(\mu, v)} &= \frac{1}{\tilde{n}} \sum_{t=1}^{\tilde{n}} \prod_{k=0}^{m-1} \widehat{C}(X_{t+k}), \\ \widehat{(v, v)} &= \frac{1}{\tilde{n}^m} \prod_{k=0}^{m-1} \left( \sum_{t=1}^{\tilde{n}} \widehat{C}(X_{t+k}) \right),\end{aligned}$$

where  $\widehat{C}(x) = \frac{1}{n} \sum_{i=1}^n \kappa(x - X_i)$  is the one-dimensional correlation integral associated with the marginal distribution. For some results concerning size and power and comparisons of those with the BDS test and the test of Granger, Maasoumi and Racine [51] see section “Multiple Bandwidth Permutation Tests”.

In fact the divergence measure  $\theta = \iint g(u)g(v) |\Psi(u, v)|^2 du dv$ , on which Pinkse [84] based his test for serial independence (see Sect. “Characteristic Functions”) is also a quadratic form (for bivariate random variables). Instead of using a U-statistics estimator of  $\theta$ , Pinkse used an estimator of a related quantity  $\vartheta$ , which in terms of the associated inner product can be expressed as:

$$\vartheta = \{[(\mu, \mu) - (\mu, v)]^2 + [(v, v) - (\mu, v)]^2\} / 2.$$

It can be verified that also  $\vartheta \geq 0$  with equality if and only if  $\vartheta = 0$ . Indeed, evidently  $\vartheta = 0$  if  $\mu = v$ , while under any alternative one cannot have both  $(\mu, v) = (\mu, \mu)$  and  $(\mu, v) = (v, v)$ , since in that case  $(\mu - v, \mu - v) = (\mu, \mu) - 2(\mu, v) + (v, v) > 0$ .

## Tests Based on Other Measures of Dependence

### Partial Sums of Data

Ashley and Patterson [2] proposed a test for independence in stock returns based on the cumulative sum  $Z_t = \sum_{j=1}^t X_j$  where  $X_j$  represents the residuals obtained after estimating an AR( $p$ ) model on returns. The idea is that if the model is appropriate, the residuals are expected to be close to i.i.d. and  $Z_t$  corresponds to the deviation of a Brownian motion on the line after  $t$  time steps. The authors proposed to test this property using the statistic  $Z^{\max} = \max\{|Z_1|, \dots, |Z_T|\}$ , assessing the statistical significance using a bootstrap method.

It was later pointed out by Corrado and Schatzberg [25] that since  $\{X_t\}$  has a zero sample mean,  $\{Z_t\}$  is ‘tied to zero’ at the endpoints ( $Z_t = Z_0 = 0$ ), and hence the reference paths used in the bootstrap should have been constructed to satisfy the same constraints. This can, for instance, be achieved by mean adjusting the bootstrap sample, or alternatively by employing a permutation method (resampling without replacement). More-

over, Corrado and Schatzberg [25] showed that after rescaling via

$$W_t = Z_t / (\sqrt{T} \hat{\sigma}_T)$$

where  $\hat{\sigma}_T$  is the sample standard deviation of  $X_j$ , the sample path of  $W_t$  for large  $T$  forms a Brownian bridge under the null hypothesis, which implies that the maximum absolute value has the same null distribution as the Kolmogorov–Smirnov (KS) test statistic.

Scaled versions of partial sums were also considered by Kulperger and Lockhart [75]. They focus on examining the conditional mean of  $Y_j := X_{j+1}$  given  $X_j$ , by studying the dependence among successive  $Y$ -values when ordered according to the ranks of the corresponding  $X$ -values. Put simply, this replaces the ordering of pairs  $(X_t, Y_t)$  in such a way that  $X$ -values are ordered increasingly, enabling the partial sums to grasp the common dependence between  $Y$ -values on  $X$ -values, rather than on time. Motivated by this, the authors propose to study the sample path of the partial sums

$$S_i = \frac{1}{\sqrt{n}} \sum_{j=1}^i (Y_{(j)} - \bar{Y}),$$

where  $Y_{(j)}$  denotes  $X_{(j)+1}$  (the successor of the observation among whose rank among the original observations is  $j$ ), and  $\bar{Y}$  is the sample mean of  $\{Y_j\}$ . The authors then propose and compare various statistics to test if the realized process  $\{S_i\}$  is a realization of a Brownian bridge, as predicted under the null hypothesis. Straightforward extensions can be obtained by taking  $Y_{(j)} = \Phi(X_{(j)+k})$  for some fixed lag  $k$ .

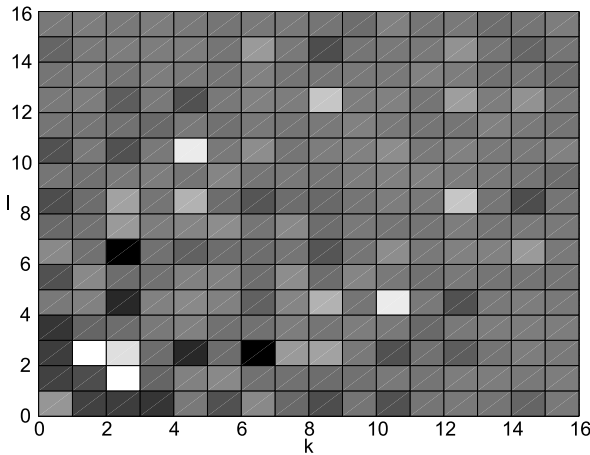
### The Spectral Density

Besides being able to test the strict random walk hypothesis (i.i.d. increments) for a financial time series such as a log-price, it is also of interest to be able to test the weaker hypothesis that increments have a constant conditional mean. A test based on the spectral density for this so-called martingale hypothesis was developed by Durlauf [41].

### The Bispectrum

As already noted by Robinson [88], one can test for serial independence against nonlinear dependence with a test for linearity rather than independence. Here and in the next subsection I briefly discuss a few examples of linearity tests.

Extending results of Subba-Rao [98], Hinich [58] used the bispectrum to detect ‘interactions between Fourier



**Nonparametric Tests for Independence, Figure 4**

Bicovariance  $c(k, \ell)$  for a bilinear process. Lighter regions correspond with larger values of the bicovariance. Series length  $n = 4000$

components'. The motivation behind the approach is that the bicorrelation, defined as

$$c(k, \ell) = E[X_t X_{t+k} X_{t+\ell}],$$

should be zero for a stationary linear Gaussian random process  $\{X_t\}$  with mean zero.

As an illustration of the structure that the bicovariance of a nonlinear time series may exhibit, Fig. 4 shows the bicovariance for the time series  $\{X_t\}$  generated by the bilinear process

$$X_t = 0.9\varepsilon_{t-1}X_{t-2} + \varepsilon_t$$

where  $\{\varepsilon_t\}$  is a sequence of independent standard normal random variables. For the simulated data I initialized the state variables at  $X_{-1} = X_0 = 0$  and discarded the first 1000 iterations.

Examining the behavior of the bicorrelation  $c(k, \ell)$  for many values of  $k$  and  $\ell$  simultaneously can be achieved in various ways, for instance by examining the bispectrum

$$B(\omega_1, \omega_2) = \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} c(k, \ell) \exp[-i(\omega_1 k + \omega_2 \ell)].$$

Hinich [58] introduced two functionals of the bispectrum that are suitable for testing Gaussianity and linearity, respectively. For applications to both model generated data and real data see, among others, [3,4,17,57]. Overall, these applications indicate that nonlinearity and non-Gaussianity play an important role in economic time series.

More recently Hinich [59] proposed a related test for serial independence of the innovations acting on a (unspecified) linear filter. This test is also based on the bispectrum, but the test statistic can be evaluated in the time

domain since it is a function of the sample bicovariance function of the residuals. Lim et al. [77] applied the test of Hinich [59] to returns of Asian market indices. If the returns would follow a GARCH process with symmetric innovations the sequence of signs thus obtained should be i.i.d. Since the results indicate nonlinear structure in the signs of Asian stock returns, the conclusion is that GARCH models with symmetric innovations are inappropriate for the returns.

Note that although this example shows how the bispectrum can be used to detect evidence against the null of (G)ARCH, the bispectrum cannot be used to test for serial independence against (G)ARCH alternatives. The reason is that the bicovariance is not able to pick up dependence in time series processes in which the conditional mean of  $X_t$  given past observations is zero, such as ARCH and GARCH processes.

If the probabilistic structure of a stationary time series process is preserved under time-reversal (i.e. when reading the series backward in time), the time series process is called time reversible. Clearly, time reversibility should hold under serial independence. Ramsey and Rothman [87,91] developed a test for time reversibility based on a sample version of the difference of bicovariances  $c(k, k) - c(0, k) = E[X_t^2 X_{t-k}] - E[X_t X_{t-k}^2]$ . This test is consistent against some forms of serial dependence, but not against all. For instance, it is not consistent against ARCH and GARCH alternatives, since these have zero bicovariance at any lag  $k$ .

Terdik and Máth [100] use the information contained in the bispectrum to test whether the best predictor is linear against quadratic alternatives. The null hypothesis being tested is linearity in mean, as opposed to a linear Gaussian random processes.

Brooks and Hinich [19,20] generalized the bispectrum approach to multivariate time series settings, in order to study nonlinear lead-lag relationships. These authors indeed found evidence for nonlinear dependence structure between various exchange rates. As noted by the authors these findings have important implications for investors who try to diversify their portfolios internationally.

### Nonlinearity in the Conditional Mean

Hjellvik and Tjøstheim [60] developed a nonparametric test for linearity based on the difference between the best linear predictor and a kernel-based nonparametric predictor. Hjellvik et al. [61] explored a variant of this approach where local polynomial predictors were used instead of kernel-based predictors.

### Bootstrap and Permutation Tests

As shown above, in many cases it is possible to use asymptotic distribution theory for test statistics in the time series context. Notable cases are those where the test statistics are U-statistics or a function thereof, as is the case for the BDS test. In some cases, however, the resulting asymptotic approximation to the finite sample null distribution may be poor. In particular this can happen when the test statistic is a degenerate or near-degenerate U-statistic under the null hypothesis. For instance, the BDS test statistic is near-degenerate under the null hypothesis if the marginal distribution is uniform.<sup>1</sup>

For practical purposes, near-degeneracy means that although the test statistic is asymptotically standard normal, it may be far from asymptotic normality even for large sample sizes. Whether a particular test statistic is (near) degenerate under the null is often not known, as it depends on the marginal distribution of the data generating process. To avoid such problems with the asymptotic approximation one can use critical values obtained by simulating a fully specified model that satisfies the i.i.d. null hypothesis. However, since the distribution of the test statistic depends on the marginal distribution, it is better to reflect this in the simulated data as well. This can be done by using bootstrap or Monte Carlo methods for assessing the statistical significance of the observed value of the test statistic of interest. The Monte Carlo procedure has the additional advantage that it produces an exact randomization test.

### Simulation

Although I do not recommend this procedure in practice, I include it for completeness. For simplicity I describe the approach here only for the one-sided case where large values of the test statistic provide evidence against the null hypothesis. The approach for two-sided tests is similar.

Suppose we wish to obtain critical values of a test statistic,  $Q_n$ , say, then this can be obtained by simulating a large number of i.i.d. time series of length  $n$ . The idea is to simulate time series satisfying the null hypothesis using a fully specified model. For instance one can simulate a large number  $B$  of i.i.d. normal time series data of the same length as the original time series, and then calculate the value of the test statistic for each of these artificial time series. The sample distribution of the  $B$  simulated test

statistics subsequently represents the null distribution, and a  $p$ -value can be obtained as  $\hat{p} = \#\{Q_n^i \geq Q_n\}/(B + 1)$ . This approach is suitable if one is willing to assume a certain marginal (normality in this case) under the null, or if the distribution of the test statistic is (at least asymptotically) independent of the (unknown) marginal distribution.

Dionísio [35] implemented a test for serial independence based on the mutual information between  $X_t$  and  $X_{t+\tau}$ . Critical values of the test statistic were obtained for individual lags  $\tau$  by simulating a large number of  $N(0, 1)$  i.i.d. time series of the appropriate length. This should provide a good approximation to the true critical values if the data are approximately normally distributed, or if the (asymptotic) distribution of the test statistic is independent of the (unknown) marginal distribution. If not, this may lead to size distortions if the data are skewed or otherwise deviating from normality. A bootstrap or a permutation test may be more appropriate in those cases.

### Bootstrap Tests

The bootstrap approach consists of resampling from the observed data with replacement. The idea is that under the null the best representation of the data generating process that we have is given by an i.i.d. process with the empirical distribution of the observed data. One of the motivating advantages of the bootstrap is that it yields an improved approximation to the finite sample distribution of test statistics relative to first-order asymptotic theory, provided that the statistics are asymptotically pivotal. For an overview of the use of the bootstrap in econometrics, I refer the interested reader to [66].

Although there are sophisticated bootstrap methods that are particularly designed for time series, for instance the block bootstrap [22,76,85], in the case of testing for serial independence the data are i.i.d. under the null, so under the null hypothesis we can bootstrap by simply drawing  $n$  values from the original time series independently with replacement. This procedure is often referred to as the naive bootstrap.

Hong and White [65] noted that the naive bootstrap does not produce a consistent procedure for their test statistic (essentially the estimator of the Kullback–Leibler divergence given in (1)) as it is degenerate under the null hypothesis of serial independence. They propose the use of a smoothed bootstrap procedure to overcome this. In the degenerate case also a permutation test may be used, as described next.

<sup>1</sup>It is exactly degenerate for i.i.d. data from the uniform distribution on the circle, i.e. the interval  $[0, a]$  with the endpoints identified [16]. Theiler [102] simulated variance of  $S = C_{m,n}(\varepsilon) - (C_{i,n}(\varepsilon))^m$  in the degenerate case, and found that it converges to 0 at the rate  $n^{-2}$  instead of the usual rate  $n^{-1}$ .

## Permutation Tests

Under the null hypothesis of serial independence the data generating process is typically known only up to an infinite dimensional nuisance parameter (the unknown marginal distribution). This prevents one from generating time series data that have the exact same distribution as the data generating process under the null hypothesis, as Barnard [5] suggested for simple null hypotheses (i.e. null hypotheses under which the distribution of the data is fully specified). Hence the problem is that the null hypothesis of serial independence is not simple but composite, with each possible marginal distribution representing another i.i.d. process. This limitation can be overcome by considering all the possible processes under the null, conditional on an observed value of a minimal sufficient statistic for the nuisance parameter. The resulting (conditional) null distribution of the test statistic can then be shown to be free of unknown parameters. The simulated data should be drawn from the same conditional distribution as the data generating process under the null, given the sufficient statistic. This procedure can be used for constructing a randomization test procedure which is exact, i.e. the type I error rate is equal to nominal level, at least in the absence of parameter estimation uncertainty. The resulting tests are referred to as Monte Carlo tests.

Since under the null hypothesis the empirical marginal distribution is a minimal sufficient statistic for the unknown marginal distribution. The conditional distribution of the observations given their empirical marginal, assigns equal probability to each of the  $n!$  possible permutations of the data. This means that every permutation of the originally observed values is equally likely. Hence an independent draw from the time series process conditional on the sufficient statistic can be obtained straightforwardly by randomly permuting the original data. The value of the test statistic for such a permuted time series is an independent draw from the conditional null distribution of the test statistic given the sufficient statistic. Although the Monte Carlo method is exact for data generated under the null hypothesis, not many investigators have studied its behavior when applied to residuals of an estimated time series model. For a general treatment of Monte Carlo testing in the presence of model parameter uncertainty, see the recent work by Dufour [37].

## Multiple Bandwidth Permutation Tests

Most of the nonparametric tests described above, such as the BDS test, require a user-specified value for the bandwidth parameter. I mentioned that the BDS test is

usually applied with bandwidth values in the range 0.5 to 1.5 standard deviations. Although these values appear to work reasonably well in numerical simulation studies with computer-generated data from known processes, there is no guarantee that this is an optimal range for the (usually unknown) alternative of most interest to the user (i.e. the possibly non-i.i.d. true process that generated the data at hand). In the different context of testing a parametric regression function against an unknown nonparametric alternative, Horowitz and Spokoiny [66] proposed tests with an adaptive bandwidth, that they showed to be rate-optimal. Although the present context is slightly different, and the details of their theorems most likely require some adaptation before they apply here, test statistics similar to the adaptive bandwidth statistics that they proposed can be easily implemented.

The idea is to calculate test statistics for many values of the bandwidth parameter,  $\varepsilon$ , say, and reject the null hypothesis if there is evidence against independence from one or more of the statistics calculated for the various bandwidths. To achieve this, an overall test statistic is required that will pick up evidence of dependence from any of the bandwidths. In Ref. [32] we proposed using the smallest  $p$ -value,  $\hat{p}(\varepsilon_i)$ , across a set of  $d$  different bandwidths  $\varepsilon_1 < \dots < \varepsilon_d$  as an overall test statistic:  $T = \inf_i \hat{p}(\varepsilon_i)$ . To establish if the value of  $T$  obtained is significant, a permutation test can be performed. I refer to this procedure as the multiple bandwidth permutation test.

Suppose that we wish to base the  $p$ -values  $\hat{p}(\varepsilon_i)$  on the permutation procedure described above, then this setup seems to require two nested permutation procedures; one global loop for replicating  $B$  values of  $T$ ,  $T_i$ ,  $i = 1, \dots, B$ , for the  $B$  different permutations of the original data, and for each of those another loop to obtain a  $p$ -values  $\hat{p}(\varepsilon_i)$  of the observed (BDS) test statistic for each bandwidth. It turns out, however, that this can be achieved much more efficiently, in a single loop across  $B$  permutations of the original data, as follows.

Let  $Q^1(\varepsilon_i)$ ,  $i = 1, \dots, b$ , denote the value of the (BDS) test statistic for the original data at the  $i$ th bandwidth,  $\varepsilon_i$ , and  $Q^k(\varepsilon_i)$ ,  $k = 2, \dots, B$ , that of the  $k$ th randomly permuted time series, then a  $p$ -value can be obtained for each bandwidth as before:  $\hat{p}^1(\varepsilon_i) = \#\{Q^s(\varepsilon_i) \geq Q^1(\varepsilon_i)\}/B$ . The superscript 1 denotes that these  $p$ -values are obtained for the original time series,  $b = 1$ . Subsequently one can obtain similar  $p$ -values for each of the permuted time series as  $\hat{p}^b(\varepsilon_i) = \#\{Q^s(\varepsilon_i) \geq Q^b(\varepsilon_i)\}/B$ . Now we are in a position to calculate the global test statistic  $T_b = \inf_i \hat{p}^b(\varepsilon_i)$  for each of the  $B$  permuted time series, including the orig-



**Nonparametric Tests for Independence, Table 1**

Observed rejection rates at nominal size 0.05 of the test of Granger, Maasoumi and Racine [51] (GMR) test and the multiple bandwidth permutation procedure for the BDS test statistic [16] and the quadratic form-based test of Diks and Panchenko [32] (DP). In the model specifications  $\{u_t\}$  represents a sequence of independent standard normal random variables. Embedding dimension  $m = 3$ , sample size 100, except for the sign process (sample size 50) and the logistic and the Hénon maps (sample size 20). Monte Carlo parameters  $B = 100$  permutations and 1000 independently realized time series from each process

Model	Specification	GMR	BDS	DP
1	$X_t = u_t$	0.05	0.05	0.05
2	$X_t = u_t + 0.8u_{t-1}^2$	0.57	0.68	0.71
3	$X_t = u_t + 0.6u_{t-1}^2 + 0.6u_{t-2}^2$	0.78	0.84	0.96
4	$X_t = u_t + 0.8u_{t-1}u_{t-2}$	0.22	0.46	0.29
5	$X_t = 0.3X_{t-1} + u_t$	0.31	0.16	0.68
6	$X_t = 0.8 X_{t-1} ^{0.5} + u_t$	0.25	0.11	0.53
7	$X_t = \text{sign}(X_{t-1}) + u_t$	0.86	0.75	0.98
8	$X_t = 0.6\varepsilon_{t-1}X_{t-2} + \varepsilon_t$	0.26	0.50	0.39
9	$X_t = \sqrt{h_t}u_t, h_t = 1 + 0.4X_{t-1}^2$	0.26	0.51	0.24
10	$X_t = \sqrt{h_t}u_t, h_t = 0.01 + 0.80h_{t-1} + 0.15X_{t-1}^2$	0.15	0.35	0.18
11	$Y_t = (-0.5 + 0.9I_{[0,\infty)}(Y_{t-1}))Y_{t-1} + \varepsilon_t$	0.34	0.06	0.87
12	$X_t = 4X_{t-1}(1 - X_{t-1}), \quad (0 < X_t < 1)$	0.95	0.71	0.90
13	$X_t = 1 + 0.3X_{t-2} - 1.4X_{t-1}^2$	0.96	0.46	0.97
14	$X_t = Z_t + \sigma u_t, \quad Z_t = 1 + 0.3Z_{t-2} - 1.4Z_{t-1}^2$	0.41	0.22	0.83

inal time series (the case  $b = 1$ ). Finally, we can establish the significance of the test statistic  $T_1$  obtained for the original time series by comparing it with the reference values  $T_2, \dots, T_B$ . Although these values need not be independent, even under the null hypothesis, they do satisfy permutation symmetry under the null hypothesis, so that each of the possible permutations of the observed values of  $T_b$  is equally likely. By the permutation symmetry of all the time series (the original and the permuted series) under the null hypothesis, and hence of the values  $T_b, b = 1, \dots, B$ , the overall  $p$ -value can still be calculated as if the values  $T_b$  were independent, i.e.  $\hat{p} = \#\{T_s \leq T_1\}/B$ . In other words, only the fact that all possible orderings of the values  $T_1, \dots, T_B$  are equally likely under the null hypothesis is needed, and not their independence.

So far I haven't discussed the possibility that ties may occur. In fact they occur with nonzero probability since  $\hat{p}^b(\varepsilon_i)$  is a discrete random variable for finite  $B$ . If ties are dealt with appropriately, however, then the above procedure leads to a test with a rejection rate under the null hypothesis equal to the nominal size (for details see [32]).

Table 1 shows the power of the multiple bandwidth procedure for the BDS test [16] and the test developed by Valentyn Panchenko and me [32] based on quadratic forms for various processes (referred to henceforth as the DP test). For comparison the power of the test of Granger Maasoumi and Racine [51] (GMR test) based on the Hellinger distance, discussed in Subsect. "Information

Theoretic Divergence Measures", are also provided. The GMR test was performed with the R software provided by the authors, which uses a bandwidth based on cross-validation.

The processes are, in order, models of type: i.i.d. normal (1), nonlinear moving average (2–4), linear autoregressive (5), nonlinear autoregressive (6), sign autoregressive (7), bilinear (8), ARCH(1) (9), GARCH(1,1) (10), threshold autoregressive (11), logistic map (12), Hénon map (13) and the Hénon map with dynamic noise (14). The multiple bandwidth permutation test was performed with 5 bandwidth values  $\varepsilon_i$  between  $\varepsilon = 0.5$  and 2.0, with a constant ratio  $\varepsilon_{i+1}/\varepsilon_i, i = 2, \dots, 4$  (hence the bandwidths are equally spaced on a logarithmic scale).

The table shows rejection rates for the i.i.d. process which are all close to the nominal size 0.05, hence there is no evidence for size distortion for any of the three tests. In terms of power (remaining processes) none of the tests does uniformly outperform the others, even within model classes such as the nonlinear moving average processes considered (process 2–4). This emphasizes again how hard it is to tell beforehand which test will perform best for an unknown alternative.

For applications I would have a slight preference for using a test that is consistent against any fixed alternative (such as the DP test), if only to hedge against the possibility of having no asymptotic power. However, as Table 1 shows, this does not guarantee a good finite sample performance in all cases.

## Future Directions

Although permutation tests have been shown to have the advantage of providing exact tests in the ideal case of data that are truly i.i.d. under the null hypothesis, more work is required to establish the properties of these (or adapted) tests in the presence of residuals of an estimated parametric model. This requires either adaptation of the permutation procedure in that setting, or analogues of the ‘nuisance parameter theorem’ for the BDS test.

Another remaining challenge is the detection of dependence within observed high-variate vector-valued time series. Estimating (functionals of) probability densities in high-dimensional spaces is notoriously difficult, since the number of observations typically required grows very fast with the number of dimensions. Due to this so-called curse of dimensionality, the kernel-based methods discussed here in practice cannot be meaningfully applied to data sets with moderate sample sizes (several thousand observations) if the dimension  $m$  exceeds 5 or 6.

Additional work is also required for the development of statistical tests for time series that do not take values in the real line, but in more general manifolds. As mentioned in the introduction, an example consists of wind direction data taking values in the interval  $[0, 2\pi]$  with the endpoints identified (i.e. on the circle). This touches upon the general problem of defining divergence measures between distributions on less familiar measurable spaces, and constructing and studying the statistical properties of their estimators.

## Bibliography

1. Aparicio FM, Escribano A (1998) Information-theoretic analysis of serial dependence and cointegration. *Stud nonlinear dyn econ* 3:119–140
2. Ashley RA, Patterson DM (1986) A nonparametric, distribution-free test for serial independence in stock returns. *J Financial Quant Anal* 21:221–227
3. Ashley RA, Patterson DM (1989) Linear versus nonlinear macroeconomics: A statistical test. *Int Econ Rev* 30:165–187
4. Ashley RA, Patterson DM, Hinich MN (1986) A diagnostic check for nonlinear serial dependence in time series fitting errors. *J Time Ser Anal* 7:165–187
5. Barnard GA (1963) Discussion of Professor Bartlett’s paper. *J Royal Stat Soc Ser B* 25:294
6. Bartels R (1982) The rank version of von Neumann’s ratio test for randomness. *J Am Stat Assoc* 77:40–46
7. Benghabrit Y, Hallin M (1992) Optimal rank-based tests against 1st-order superdiagonal bilinear dependence. *J Stat Plan Inference* 32:45–61
8. Bera AK, Robinson PM (1989) Tests for serial dependence and other specification analysis in models of markets in equilibrium. *J Bus Econ Stat* 7:343–352
9. Beran J (1992) A goodness-of-fit test for time-series with long-range dependence. *J Royal Stat Soc Ser B* 54:749–760
10. Blum JR, Kiefer J, Rosenblatt M (1961) Distribution free tests of independence based on sample distribution functions. *Ann Math Stat* 32:485–498
11. Bollerslev T (1986) Generalized autoregressive heteroskedasticity. *J Econometrics* 31:307–327
12. Booth GG, Martikainen T (1994) Nonlinear dependence in Finnish stock returns. *Eur J Oper Res* 74:273–283
13. Box GEP, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J Am Stat Assoc* 332:1509–1526
14. Bradley R (1986) Basic properties of strong mixing conditions. In: Eberlein E, Taqqu MS (eds) *Dependence in Probability and Statistics*. Birkhäuser, Basel
15. Brock WA, Dechert WD, Scheinkman JA (1987) A test for independence based on the correlation dimension. Working paper 8702. University of Wisconsin, Madison
16. Brock WA, Dechert WD, Scheinkman JA, LeBaron B (1996) A test for independence based on the correlation dimension. *Econometric Rev* 15:197–235
17. Brockett PL, Hinich MD, Patterson D (1988) Bispectral based tests for the detection of Gaussianity and linearity in time series. *J Am Stat Assoc* 83:657–664
18. Brooks C, Heravi SM (1999) The effect of (mis-specified) GARCH filters on the finite sample distribution of the BDS test. *Comput Econ* 13:147–162
19. Brooks C, Hinich MJ (1999) Cross-correlations and cross-bicorrelations in Sterling exchange rates. *J Empir Finance* 6:385–404
20. Brooks C, Hinich MJ (2001) Bicorrelations and cross-bicorrelations as non-linearity tests and tools for exchange rate forecasting. *J Forecast* 20:181–196
21. Caporale GM, Ntantamis C, Pantelidis T, Pittis N (2005) The BDS test as a test for the adequacy of a GARCH(1,1) specification: A Monte Carlo study. *J Financial Econometric* 3:282–309
22. Carlstein E (1984) The use of sub-series methods for estimating the variance of a general statistic from a stationary time series. *Ann Stat* 14:1171–1179
23. Carlstein E (1988) Degenerate U-statistics based on non-independent observations. *Calcutta Stat Assoc Bull* 37:55–65
24. Chan NH, Tran LT (1992) Nonparametric tests for serial independence. *J Time Ser Anal* 13:19–28
25. Corrado CJ, Schatzberg J (1990) A nonparametric, distribution-free test for serial independence in stock returns: A correction. *J Financial Quant Anal* 25:411–415
26. Csörgő S (1985) Testing for independence by the empirical characteristic function. *J Multivar Anal* 16:290–299
27. Debnath L, Mikusiński P (2005) *Introduction to Hilbert Spaces With Applications*, 3rd edn. Elsevier Academic Press, Burlington
28. Delgado M, Mora J (2000) A nonparametric test for serial independence of regression errors. *Biometrika* 87:228–234
29. Delgado MA (1996) Testing serial independence using the sample distribution function. *J Time Ser Anal* 11:271–285
30. Denker M, Keller G (1983) On U-statistics and v. Mises’ statistics for weakly dependent processes. *Z Wahrscheinlichkeitstheorie verwandte Geb* 64:505–522
31. Denker M, Keller G (1986) Rigorous statistical procedures for data from dynamical systems. *J Stat Phys* 44:67–93
32. Diks C, Panchenko V (2007) Nonparametric tests for se-

- rial independence based on quadratic forms. *Statistica Sin* 17:81–97
33. Diks C, Panchenko V (2008) Rank-based entropy tests for serial independence. *Stud Nonlinear Dyn Econom* 12(1):art.2:0–19
  34. Diks C, Tong H (1999) A test for symmetries of multivariate probability distributions. *Biometrika* 86:605–614
  35. Dionísio A, Menezes R, Mendes DA (2006) Entropy-based independence test. *Nonlinear Dyn* 44:351–357
  36. Dufour JM (1981) Rank tests for serial dependence. *J Time Ser Anal* 2:117–128
  37. Dufour JM (2006) Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and non-standard asymptotics. *J Econom* 133:443–477
  38. Durbin J, Watson GS (1950) Testing for serial correlation in least-squares regression, I. *Biometrika* 37:409–428
  39. Durbin J, Watson GS (1951) Testing for serial correlation in least-squares regression, II. *Biometrika* 38:159–177
  40. Durbin J, Watson GS (1971) Testing for serial correlation in least-squares regression, III. *Biometrika* 58:1–19
  41. Durlauf S (1991) Spectral based testing of the martingale hypothesis. *J Econometrics* 50:355–376
  42. Engle R (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50:987–1007
  43. Ferguson TS, Genest C, Hallin M (2000) Kendall's tau for serial dependence. *Can J Stat* 28:587–604
  44. Fernandes M, Neri B (2008) Nonparametric entropy-based tests of independence between stochastic processes. *Econometric Reviews*; Forthcoming
  45. Genest C, Quessy JF, Rémillard B (2002) Tests of serial independence based on Kendall's process. *Can J Stat* 30:1–21
  46. Genest C, Rémillard B (2004) Tests of independence and randomness based on the empirical copula process. *Test* 13:335–369
  47. Genest C, Verret F (2005) Locally most powerful rank tests of independence for copula models. *Nonparametric Stat* 17:521–539
  48. Genest C, Ghouli K, Rémillard B (2007) Rank-based extensions of the Brock, Dechert, and Scheinkman test. *J Am Stat Assoc* 102:1363–1376
  49. Ghouli K, Kulperger RJ, Rémillard B (2001) A nonparametric test of serial independence for time series and residuals. *J Multivar Anal* 79:191–218
  50. Granger C, Lin JL (2001) Using the mutual information coefficient to identify lags in nonlinear models. *J Time Ser Anal* 15:371–384
  51. Granger CW, Maasoumi E, Racine J (2004) A dependence metric for possibly nonlinear processes. *J Time Ser Anal* 25:649–669
  52. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. *Physica D* 9:189–208
  53. Grassberger P, Schreiber T, Schaffrath C (1991) Nonlinear time sequence analysis. *Int J Bifurc Chaos* 1:521–547
  54. Hallin M, Ingenbleek J-F, Puri ML (1985) Linear serial rank tests for randomness against ARMA alternatives. *Ann Stat* 13:1156–1181
  55. Hallin M, Mélard G (1988) Rank-based tests for randomness against first-order serial dependence. *J Am Stat Assoc* 83:1117–1128
  56. Hannan EJ (1957) Testing for serial correlation in least squares regression. *Biometrika* 44:57–66
  57. Hinich M, Patterson D (1985) Evidence of nonlinearity in stock returns. *J Bus Econ Stat* 3:69–77
  58. Hinich MJ (1982) Testing for Gaussianity and linearity of a stationary time series. *J Time Ser Anal* 3:169–176
  59. Hinich MJ (1996) Testing for dependence in the input to a linear time series model. *J Nonparametric Stat* 8:205–221
  60. Hjellvik V, Tjøstheim D (1995) Nonparametric tests of linearity for time series. *Biometrika* 82:351–368
  61. Hjellvik V, Yao Q, Tjøstheim D (1998) Linearity testing using polynomial approximation. *J Stat Plan Inference* 68:295–321
  62. Hoeffding W (1948) A non-parametric test of independence. *Ann Math Stat* 19:546–557
  63. Hong Y (1999) Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *J Am Stat Assoc* 94:1201–1220
  64. Hong Y (2000) Generalized spectral tests for serial dependence. *J Royal Stat Soc Ser B* 62:557–574
  65. Hong Y, White H (2005) Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* 73:837–901
  66. Horowitz JL (2001) The bootstrap. In: Heckman JJ, Leamer EE (eds) *Handbook of Econometrics*, vol 5. Elsevier, Amsterdam, pp 3159–3228
  67. Horowitz JL, Spokoiny VG (2001) An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* 69:599–631
  68. Joe H (1989) Relative entropy measures of multivariate dependence. *J Am Stat Assoc* 84:157–164
  69. Joe H (1990) Multivariate concordance. *J Multivar Anal* 35:12–30
  70. Johnson D, McLelland R (1997) Nonparametric tests for the independence of regressors and disturbances as specification tests. *Rev Econ Stat* 79:335–340
  71. Johnson D, McLelland R (1998) A general dependence test and applications. *J Appl Econometrics* 13:627–644
  72. Kallenberg WCM, Ledwina T (1999) Data driven rank tests for independence. *J Am Stat Assoc* 94:285–301
  73. Kendall MG (1938) A new measure of rank correlation. *Biometrika* 30:81–93
  74. Kočenda E, Briatka L' (2005) Optimal range for the IID test based on integration across the correlation integral. *Econometric Rev* 24:265–296
  75. Kulperger RJ, Lockhart RA (1998) Tests of independence in time series. *J Time Ser Anal* 1998:165–185
  76. Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. *Ann Stat* 17:1217–1241
  77. Lim KP, Hinich MJ, Liew VKS (2005) Statistical inadequacy of GARCH models for Asian stock markets: Evidence and implications. *J Emerg Mark Finance* 4:263–279
  78. Lima P De (1996) Nuisance parameter free properties of correlation integral based statistics. *Econometric Rev* 15:237–259
  79. Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. *Biometrika* 65:297–302
  80. Lo AW (2000) Finance: A selective survey. *J Am Stat Assoc* 95:629–635
  81. Maasoumi E (2002) Entropy and predictability of stock market returns. *J Econometrics* 107:291–312
  82. McLeod AI, Li WK (1983) Diagnostic checking ARMA time series models using squared-residual autocorrelations. *J Time*

Ser Anal 4:269–273

83. Von Neumann J (1941) Distribution of the ratio of the mean square successive difference to the variance. *Ann Math Stat* 12:367–395
84. Pinkse J (1998) A consistent nonparametric test for serial independence. *J Econometrics* 84:205–231
85. Politis DN, Romano JP (1994) The stationary bootstrap. *J Am Stat Assoc* 89:1303–1313
86. Racine J, Maasoumi E (2007) A versatile and robust metric entropy test for time-irreversibility, and other hypotheses. *J Econometrics* 138:547–567
87. Ramsey JB, Rothman P (1990) Time irreversibility of stationary time series: estimators and test statistics. Unpublished manuscript, Department of Economics, New York University and University of Delaware
88. Robinson PM (1991) Consistent nonparametric entropy-based testing. *Rev Econ Stud* 58:437–453
89. Rosenblatt M (1975) A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann Stat* 3:1–14
90. Rosenblatt M, Wahlen BE (1992) A nonparametric measure of independence under a hypothesis of independent components. *Stat Probab Lett* 15:245–252
91. Rothman P (1992) The comparative power of the TR test against simple threshold models. *J Appl Econometrics* 7:S187–S195
92. Scaillet O (2005) A Kolmogorov–Smirnov type test for positive quadrant dependence. *Can J Stat* 33:415–427
93. Serfling RJ (1980) *Approximation Theorems of Mathematical Statistics*. Wiley, New York
94. Silverman BW (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York
95. Skaug HJ, Tjøstheim D (1993a) A nonparametric test for serial independence based on the empirical distribution function. *Biometrika* 80:591–602
96. Skaug HJ, Tjøstheim D (1993b) Nonparametric tests of serial independence. In: Subba Rao T (ed) *Developments in Time Series Analysis: the M. B. Priestley Birthday Volume*. Wiley, New York
97. Spearman C (1904) The proof and measurement of association between two things. *Am J Psychol* 15:72–101
98. Subba Rao T, Gabr MM (1980) A test for linearity of stationary time series. *J Time Ser Anal* 1:145–158
99. Takens F (1981) Detecting strange attractors in turbulence. In: Rand DA, Young LS (eds) *Dynamical Systems and Turbulence*, Warwick 1980. (Lecture Notes in Mathematics), vol 898. Springer, Berlin, pp 366–381
100. Terdik G, Máth J (1998) A new test of linearity of time series based on the bispectrum. *J Time Ser Anal* 19:737–753
101. Theil H, Nagar AL (1961) Testing the independence of regression disturbances. *J Am Stat Assoc* 56:793–806
102. Theiler J (1990) Statistical precision of dimension estimators. *Phys Rev A* 41:3038–3051
103. Tjøstheim D (1996) Measures of dependence and tests of independence. *Statistics* 28:249–284
104. Tong H (1990) *Non-linear Time Series: A Dynamical Systems Approach*. Clarendon Press, Oxford
105. Tsallis C (1998) Generalized entropy-based criterion for consistent testing. *Phys Rev E* 58:1442–1445
106. Wolff RC (1994) Independence in time series: another look at the BDS test. *Philos Trans Royal Soc Ser A* 348:383–395

## Nonsmooth Analysis in Systems and Control Theory

FRANCIS CLARKE

Institut universitaire de France et Université de Lyon,  
Lyon, France

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Elements of Nonsmooth Analysis](#)

[Necessary Conditions in Optimal Control](#)

[Verification Functions](#)

[Dynamic Programming and Viscosity Solutions](#)

[Lyapunov Functions](#)

[Stabilizing Feedback](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Generalized gradients and subgradients** These terms refer to various set-valued replacements for the usual derivative which are used in developing differential calculus for functions which are not differentiable in the classical sense. The subject itself is known as *nonsmooth analysis*. One of the best-known theories of this type is that of *generalized gradients*. Another basic construct is the *subgradient*, of which there are several variants. The approach also features *generalized tangent and normal vectors* which apply to sets which are not classical manifolds. The article contains a summary of the essential definitions.

**Pontryagin Maximum Principle** The main theorem on necessary conditions in optimal control was developed in the 1950s by the Russian mathematician L. Pontryagin and his associates. The Maximum Principle unifies and extends to the control setting the classical necessary conditions of Euler and Weierstrass from the calculus of variations, as well as the transversality conditions. There have been numerous extensions since then, as the need to consider new types of problems continues to arise.

**Verification functions** In attempting to prove that a certain control is indeed the solution to a given optimal control problem, one important approach hinges upon exhibiting a function having certain properties implying the optimality of the given control. Such a function is termed a verification function. The approach



becomes widely applicable if one allows nonsmooth verification functions.

**Dynamic programming** A well-known technique in dynamic problems of optimization is to solve (in a discrete context) a backwards recursion for a certain value function related to the problem. This technique, which was developed notably by Bellman, can be applied in particular to optimal control problems. In the continuous setting, the recursion corresponds to the *Hamilton–Jacobi Equation*. This partial differential equation does not generally admit smooth classical solutions. The theory of *viscosity solutions* uses subgradients to define generalized solutions, and obtains their existence and uniqueness.

**Lyapunov function** In the classical theory of ordinary differential equations, global asymptotic stability is most often verified by exhibiting a *Lyapunov function*, a function along which trajectories decrease. In that setting, the existence of a smooth Lyapunov function is both necessary and sufficient for stability. The Lyapunov function concept can be extended to control systems, but in that case it turns out that nonsmooth functions are essential. These generalized *control Lyapunov functions* play an important role in designing optimal or stabilizing feedback.

### Definition of the Subject

The term *nonsmooth analysis* refers to the body of theory which develops differential calculus for functions which are not differentiable in the usual sense, and for sets which are not classical smooth manifolds. There are several different (but related) approaches to doing this. Among the better-known constructs of the theory are the following: generalized gradients and Jacobians, proximal subgradients, subdifferentials, generalized directional (or Dini) derivatives, together with various associated tangent and normal cones. Nonsmooth analysis is a subject in itself, within the larger mathematical field of differential (variational) analysis or functional analysis, but it has also played an increasingly important role in several areas of application, notably in optimization, calculus of variations, differential equations, mechanics, and control theory. Among those who have participated in its development (in addition to the author) are J. Borwein, A. D. Ioffe, B. Morukhovich, R. T. Rockafellar, and R. B. Vinter, but many more have contributed as well.

In the case of control theory, the need for nonsmooth analysis first came to light in connection with finding proofs of necessary conditions for optimal control, notably in connection with the Pontryagin Maximum Prin-

ciple. This necessity holds even for problems which are expressed entirely in terms of smooth data. Subsequently, it became clear that problems with intrinsically nonsmooth data arise naturally in a variety of optimal control settings. Generally, nonsmooth analysis enters the picture as soon as we consider problems which are truly nonlinear or non-linearizable, whether for deriving or expressing necessary conditions, in applying sufficient conditions, or in studying the sensitivity of the problem.

The need to consider nonsmoothness in the case of stabilizing (as opposed to optimal) control has come to light more recently. It appears in particular that in the analysis of truly nonlinear control systems, the consideration of nonsmooth Lyapunov functions and discontinuous feedbacks becomes unavoidable.

### Introduction

The basic object in the control theory of ordinary differential equations is the system

$$\dot{x}(t) = f(x(t), u(t)) \text{ a.e.}, \quad 0 \leq t \leq T, \quad (1)$$

where the (measurable) *control function*  $u(\cdot)$  is chosen subject to the constraint

$$u(t) \in U \text{ a.e.}, \quad (2)$$

where  $U$  is a given set in an Euclidean space, and where the ensuing *state*  $x(\cdot)$  (a function with values in  $\mathbb{R}^n$ ) is subject to certain conditions, including most often an initial one of the form  $x(0) = x_0$ , and perhaps other constraints, either throughout the interval (pointwise) or at the terminal time. A control function  $u$  of this type is referred to as an *open loop control*. This indirect control of  $x(\cdot)$  via the choice of  $u(\cdot)$  is to be exercised for a purpose, of which there are two principal sorts:

- *positional*:  $x(t)$  is to remain in a given set in  $\mathbb{R}^n$ , or approach that set;
- *optimal*:  $x(\cdot)$ , together with  $u(\cdot)$ , is to minimize a given functional.

The second of these criteria follows directly in the tradition of the calculus of variations, and gives rise to the subject of *optimal control*, in which the dominant issues are those of optimization: necessary conditions for optimality, sufficient conditions, regularity of the optimal control, sensitivity. We discuss below the role of nonsmooth analysis in optimal control; this was the setting of many of the earliest applications.

In contrast, a prototypical control problem of purely positional sort would be the following:

Find a control  $u(\cdot)$  such that  $x(\cdot)$  goes to 0.



This rather vaguely worded goal is often more precisely expressed as that of finding a *stabilizing feedback control*  $k(x)$ ; that is, a function  $k$  with values in  $U$  (this is often referred to as a closed-loop control) such that all solutions of the differential equation

$$\dot{x}(t) = f(x(t), k(x(t))), \quad x(0) = \alpha \quad (3)$$

converge to 0 (for all values of  $\alpha$ ) in a suitable sense.

The most common approach to designing the required stabilizing feedback uses the technique that is central to most of applied mathematics: *linearization*. In this case, one examines the linearized system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

where

$$A := f_x(0, 0), \quad B := f_u(0, 0).$$

If the linearized system satisfies certain controllability properties, then classical linear systems theory provides well-known and powerful tools for designing (linear) feedbacks that stabilize the linearized system. Under further mild hypotheses, this yields a feedback that stabilizes the original nonlinear system (1) *locally*; that is, for initial values  $x(0)$  sufficiently near 0.

This approach has been feasible in a large number of cases, and in fact it underlies the very successful role that control theory has played in a great variety of applications. Still, linearization does require that a certain number of conditions be met:

- The function  $f$  must be smooth (differentiable) so that the linear system can be constructed;
- The linear system must be a ‘nondegenerate’ approximation of the nonlinear one (that is, it must be controllable);
- The control set  $U$  must contain a neighborhood of 0, so that near 0 the choice of controls is unconstrained, and all state values near 0 must be acceptable (no state constraints);
- Both  $x$  and  $u$  must remain small so that the linear approximation remains relevant (in dealing with errors or perturbations, the feedback is operative only when they are sufficiently small).

It is not hard to envisage situations in which the last two conditions fail, and indeed such challenging problems are beginning to arise increasingly often. The first condition fails for simple problems involving electrical circuits in which a diode is present, for example (see [14]). A famous (smooth) mechanical system for which the second condition fails is the *nonholonomic integrator*, a term which

refers to the following system, which is linear (separately) in the state and in the control variables:

$$\begin{cases} \dot{x}_1 = u_1 \\ \dot{x}_2 = u_2 \\ \dot{x}_3 = x_1 u_2 - x_2 u_1 \end{cases} \quad \|(u_1, u_2)\| \leq 1.$$

(Thus  $n = 3$  here, and  $U$  is the closed unit ball in  $\mathbb{R}^2$ .) Here, the linearized system is degenerate, since its third component is  $\dot{x}_3 = 0$ . As discussed below, there is in fact no continuous feedback law  $u = k(x)$  which will stabilize this system (even locally about the origin), but certain discontinuous stabilizing feedbacks do exist.

This illustrates the moral that when linearization is not applicable to a given nonlinear system (for whatever reason), nonsmoothness generally arises. (This has been observed in other contexts in recent decades: catastrophes, chaos, fractals.) Consider for example the issue of whether a (control) Lyapunov function exists. This refers to a pair  $(V, W)$  of positive definite functions satisfying notably the following *Infinitesimal Decrease* condition:

$$\inf_{u \in U} \langle \nabla V(x), f(x, u) \rangle \leq -W(x) \quad x \neq 0.$$

The existence of such (smooth) functions implies that the underlying control system is *globally asymptotically controllable* (GAC), which is a necessary condition for the existence of a stabilizing feedback (and also sufficient, as recently proved [21]). In fact, exhibiting a Lyapunov function is the principal technique for proving that a given system is GAC; the function  $V$  in question then goes on to play a role in designing the stabilizing feedback.

It turns out, however, that even smooth systems that are GAC need not admit a smooth Lyapunov function. (The nonholonomic integrator is an example of this phenomenon.) But if one extends in a suitable way the concept of Lyapunov function to nonsmooth functions, then the existence of a Lyapunov function becomes a necessary and sufficient condition for a given system to be GAC. One such extension involves replacing the gradient that appears in the Infinitesimal Decrease condition above by elements of the *proximal subdifferential* (see below). How to use such extended Lyapunov functions to design a stabilizing feedback is a nontrivial topic that has only recently been successfully addressed.

In the next section we define a few basic constructs of nonsmooth analysis; knowledge of these is sufficient for interpreting the statements of the results discussed in this article.

## Elements of Nonsmooth Analysis

This section summarizes some basic notions in nonsmooth analysis, in fact a minimum so that the statements of the results to come can be understood. This minimum corresponds to three types of set-valued generalized derivative (generalized gradients, proximal subgradients, limiting subgradients), together with a notion of normal vector applicable to any closed (not necessarily smooth or convex) set. (See [24] for a thorough treatment and detailed references.)

### Generalized Gradients

For *smooth* real-valued functions  $f$  on  $\mathbb{R}^n$  we have a well-known formula linking the usual directional derivative to the gradient:

$$f'(x; v) := \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t} = \langle \nabla f(x), v \rangle.$$

We can extend this pattern to *Lipschitz* functions:  $f$  is said to be Lipschitz on a set  $S$  if there is a constant  $K$  such that  $|f(y) - f(z)| \leq K\|y - z\|$  whenever  $y$  and  $z$  belong to  $S$ . A function  $f$  that is Lipschitz in a neighborhood of a point  $x$  is not necessarily differentiable at  $x$ , but we can define a *generalized directional derivative* as follows:

$$f^\circ(x; v) := \limsup_{y \rightarrow x, t \downarrow 0} \frac{f(y + tv) - f(y)}{t}.$$

Having done so, we proceed to define the *generalized gradient*:

$$\partial_C f(x) := \{ \zeta \in \mathbb{R}^n : f^\circ(x; v) \geq \langle \zeta, v \rangle \quad \forall v \in X \}.$$

It turns out that  $\partial_C f(x)$  is a compact convex nonempty set that has a calculus reminiscent of the usual differential calculus; for example, we have  $\partial(-f)(x) = -\partial f(x)$ . We also have  $\partial(f + g)(x) \subset \partial f(x) + \partial g(x)$ ; note that (as is often the case) this is an inclusion, not an equation. There is also a useful analogue of the Mean Value Theorem, and other familiar results. In addition, though, there are new formulas having no smooth counterpart, such as one for the generalized gradient of the pointwise maximum of locally Lipschitz functions ( $\partial\{\max_{1 \leq i \leq n} f_i(x)\} \subset \dots$ ).

A very useful fact for actually calculating  $\partial_C f(x)$  is the following Gradient Formula:

$$\partial_C f(x) = \text{co} \left\{ \lim_{i \rightarrow \infty} \nabla f(x_i) : x_i \rightarrow x, x_i \notin \Gamma \right\},$$

where  $\Gamma$  is *any* set of measure zero containing the local points of nondifferentiability of  $f$ . Thus the generalized gradient is ‘blind to sets of measure zero’.

Generalized gradients and their calculus were defined by Clarke [12] in 1973. The theory can be developed on any Banach space; the infinite-dimensional context is essential in certain control applications, but for our present purposes it suffices to limit attention to functions defined on  $\mathbb{R}^n$ . There is in addition a corresponding theory of tangent and normal vectors to arbitrary closed sets; we give some elements of this below.

### Proximal Subgradients

We now present a different approach to developing nonsmooth calculus, one that uses the notion of *proximal subgradient*. Let  $f: \mathbb{R}^n \rightarrow (-\infty, \infty]$  be a given function (note that the value  $+\infty$  is admitted here), and let  $x$  be a point where  $f(x)$  is finite. A vector  $\zeta$  in  $\mathbb{R}^n$  is said to be a proximal subgradient of  $f$  at  $x$  provided that there exist a neighborhood  $\Omega$  of  $x$  and a number  $\sigma \geq 0$  such that

$$f(y) \geq f(x) + \langle \zeta, y - x \rangle - \sigma \|y - x\|^2 \quad \forall y \in \Omega.$$

Thus the existence of a proximal subgradient  $\zeta$  at  $x$  corresponds to the possibility of approximating  $f$  from below (thus in a *one-sided* manner) by a function whose graph is a parabola. The point  $(x, f(x))$  is a contact point between the graph of  $f$  and the parabola, and  $\zeta$  is the slope of the parabola at that point. Compare this with the usual derivative, in which the graph of  $f$  is approximated (in a two-sided way) by an affine function.

The set of proximal subgradients at  $x$  (which may be empty, and which is not necessarily closed, open, or bounded but which is convex) is denoted  $\partial_P f(x)$ , and is referred to as the *proximal subdifferential*. If  $f$  is differentiable at  $x$ , then we have  $\partial_P f(x) \subset \{f'(x)\}$ ; equality holds if  $f$  is of class  $C^2$  at  $x$ .

As a guide to understanding, the reader may wish to carry out the following exercise (in dimension  $n = 1$ ): the proximal subdifferential at 0 of the function  $f_1(x) := -|x|$  is empty, while that of  $f_2(x) := |x|$  is the interval  $[-1, 1]$ .

The *proximal density theorem* asserts that  $\partial_P f(x)$  is nonempty for all  $x$  in a dense subset of

$$\text{dom } f := \{x : f(x) < \infty\}.$$

Although it can be empty at many points, the proximal subgradient admits a very complete calculus for the class of lower semicontinuous functions: all the usual calculus rules that the reader knows (and more) have their counterpart in terms of  $\partial_P f$ . Let us quote for example Ioffe’s *fuzzy sum rule*: if  $\zeta \in \partial_P(f + g)(x)$ , then for any  $\epsilon > 0$  there exist  $x'$  and  $x''$  within  $\epsilon$  of  $x$ , together with points

$\zeta' \in \partial_P f(x')$  and  $\zeta'' \in \partial_P g(x'')$  such that

$$\zeta \in \zeta' + \zeta'' + \epsilon B.$$

### The Limiting Subdifferential

A sequential closure operation applied to  $\partial_P f$  gives rise to the *limiting subdifferential*, useful for stating results:

$$\partial_L f(x) := \{\lim \zeta_i : \zeta_i \in \partial_P f(x_i), x_i \rightarrow x, f(x_i) \rightarrow f(x)\}.$$

If  $f$  is Lipschitz near  $x$ , then  $\partial_L f(x)$  is nonempty, and (for any lower semicontinuous  $g$  finite at  $x$ ) we have

$$\partial_L(f + g)(x) \subset \partial_L f(x) + \partial_L g(x).$$

When the function  $f$  is Lipschitz near  $x$ , then both the approaches given above (generalized gradients, proximal subdifferential) apply, and the corresponding constructs are related as follows:

$$\partial_C f(x) = \text{co } \partial_L f(x).$$

### Normal Vectors

Given a nonempty closed subset  $S$  of  $\mathbb{R}^n$  and a point  $x$  in  $S$ , we say that  $\zeta \in X$  is a *proximal normal* (vector) to  $S$  at  $x$  if there exists  $\sigma \geq 0$  such that

$$\langle \zeta, x' - x \rangle \leq \sigma \|x' - x\|^2 \quad \forall x' \in S.$$

(This is the *proximal normal inequality*.) The set (convex cone) of such  $\zeta$ , which always contains 0, is denoted  $N_S^P(x)$  and referred to as the proximal normal cone. We apply to  $N_S^P(x)$  a sequential closure operation in order to obtain the *limiting normal cone*:

$$N_S^L(x) := \{\lim \zeta_i : \zeta_i \in N_S^P(x_i), x_i \rightarrow x, x_i \in S\}.$$

These geometric notions are consistent with the analytical ones, as illustrated by the formulas

$$\partial_P I_S(x) = N_S^P(x), \quad \partial_L I_S(x) = N_S^L(x),$$

where  $I_S$  denotes the *indicator* of the set  $S$ : the function which equals 0 on  $S$  and  $+\infty$  elsewhere. They are also consistent with the more traditional ones: When  $S$  is either a convex set, a smooth manifold, or a manifold with boundary, then both  $N_S^P(x)$  and  $N_S^L(x)$  coincide with the usual normal vectors (a cone, space, or half-space respectively).

### Viscosity Subdifferentials

We remark that the *viscosity subdifferential* of  $f$  at  $x$  (commonly employed in pde's, but not used in this article) corresponds to the set of  $\zeta$  for which we have

$$f(y) \geq f(x) + \langle \zeta, y - x \rangle + \theta(|y - x|),$$

where  $\theta$  is a function such that  $\lim_{t \downarrow 0} \theta(t)/t = 0$ . This gives a potentially larger set than  $\partial_P f(x)$ ; however, the viscosity subdifferential satisfies the same (fuzzy) calculus rules as does  $\partial_P f$ . In addition, the sequential closure operation described above gives rise to the *same* limiting subdifferential  $\partial_L f(x)$  (see [24]). In most cases, therefore, it is equivalent to work with viscosity or proximal subgradients.

### Necessary Conditions in Optimal Control

The central theorem on necessary conditions for optimal control is the Pontryagin Maximum Principle. (The literature on necessary conditions in optimal control is now very extensive; we cite some standard references in the bibliography.) Even in the somewhat special (by current standards) smooth context in which it was first proved [47], an element of generalized differentiability was required. With the subsequent increase in both the generality of the model and weakening of the hypotheses (all driven by real applications), the need for nonsmooth analysis is all the greater, even for the very statement of the result.

We give here just one example, a broadly applicable *hybrid maximum principle* taken from [11], in order to highlight the essential role played by nonsmooth analysis as well as the resulting versatility of the results obtained.

### The Problem and Basic Hypotheses

We consider the minimization of the functional

$$\ell(x(a), x(b)) + \int_a^b F(t, x(t), u(t)) dt$$

subject to the boundary conditions  $(x(a), x(b)) \in S$  and the standard control dynamics

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e.}$$

The minimization takes place with respect to (absolutely continuous) state arcs  $x$  and measurable control functions  $u: [a, b] \rightarrow \mathbb{R}^m$ . Note that no explicit constraints are placed upon  $u(t)$ ; if such constraints exist, they are accounted for by assigning to the extended-valued integrand  $F$  the value  $+\infty$  whenever the constraints are violated. It is part of our intention here to demonstrate the

utility of indirectly taking account of constraints in this way. As basic hypotheses, we assume that  $F$  is measurable and lower semicontinuous in  $(x, u)$ ;  $\ell$  is taken to be locally Lipschitz and  $S$  closed. For the theorem below, we require that  $f$  be measurable in  $t$  and continuously differentiable in  $(x, u)$  (but see the remarks for a weakening of the smoothness).

### The Growth Conditions

The principal hypothesis here is that  $f$  and  $F$  satisfy the following: for every bounded subset  $X$  of  $\mathbb{R}^n$ , there exist a constant  $c$  and a summable function  $d$  such that, for almost every  $t$ , for every  $(x, u) \in \text{dom } F(t, \cdot, \cdot)$  with  $x \in X$ , we have

$$\|D_x f(t, x, u)\| \leq c \{|f(t, x, u)| + F(t, x, u)\} + d(t),$$

and for all  $(\zeta, \psi)$  in  $\partial_P F(t, x, u)$  (if any) we have

$$\frac{|\zeta| (1 + \|D_u f(t, x, u)\|)}{1 + |\psi|} \leq c \{|f(t, x, u)| + F(t, x, u)\} + d(t).$$

In the following,  $(*)$  denotes evaluation at  $(x_*(t), u_*(t))$ .

**Theorem 1** *Let the control function  $u_*$  give rise to an arc  $x_*$  which is a local minimum for the problem above. Then there exist an arc  $p$  on  $[a, b]$  such that the following transversality condition holds*

$$(p(a), -p(b)) \in \partial_L \ell(x_*(a), x_*(b)) + N_S^L(x_*(a), x_*(b)),$$

*and such that  $p$  satisfies the adjoint inclusion:  $\dot{p}(t)$  belongs almost everywhere to the set*

$$\text{co}\{\omega: (\omega + D_x^* f(t, *)p(t), D_u^* f(t, *)p(t)) \in \partial_L F(t, *)\},$$

*as well as the maximum condition: for almost every  $t$ , for every  $u$  in  $\text{dom } F(t, x_*(t), \cdot)$ , one has*

$$\begin{aligned} &\langle p(t), f(t, x_*(t), u) \rangle - F(t, x_*(t), u) \\ &\leq \langle p(t), f(t, *) \rangle - F(t, *). \end{aligned}$$

We proceed to make a few remarks on this theorem, beginning with the fact that it can fail in the absence of the growth condition, even for smooth problems of standard type [30]. The statement of the theorem is not complete, since in general the necessary conditions may hold only in *abnormal* form; we do not discuss this technical point here for reasons of economy. The phrase ‘local minimum’ (which we have also not defined) can be interpreted very generally, see [11]. The ‘maximum condition’ above is of

course what leads to the name ‘maximum principle’, while the exotic-looking adjoint inclusion reduces to more familiar conditions in a variety of special cases (see below). The transversality condition illustrates well the utility of couching in nonsmooth analysis terms the conclusion, since the given form encapsulates simultaneously a wide variety of conclusions obtainable in special cases. To give but one example, consider an optimal control problem in which  $x(a)$  is prescribed,  $x(b)$  does not appear in the cost, but is subject to an inequality constraint  $g(x(b)) \leq 0$  for a certain smooth scalar function  $g$ . Then the transversality condition of the theorem (interpreted via standard facts in nonsmooth analysis) asserts that  $p(b)$  is of the form  $\lambda \nabla g(x_*(b))$  for some nonpositive number  $\lambda$  (a Lagrange multiplier).

To illustrate the versatility of the theorem, we look at some special cases.

### The Standard Problem

The first case we examine is that in which for each  $t$ ,

$$F(t, x, u) = I_{U(t)}(u),$$

the indicator function which takes the value 0 when  $u \in U(t)$  and  $+\infty$  otherwise. This simply corresponds to imposing the condition  $u(t) \in U(t)$  on the admissible controls  $u$ ; this is the *Mayer form* of the problem (no integral term in the cost, obtained by reformulation if necessary). Note that in this case the second growth condition is trivially satisfied (since  $\zeta = 0$ ). The first growth condition is active only on  $U(t)$ , and certainly holds if  $f$  is smooth in  $(t, x, u)$  and  $U(t)$  is uniformly bounded. The hybrid adjoint inclusion immediately implies the standard adjoint equation

$$-\dot{p}(t) = D_x^* f(t, x_*(t), u_*(t)) p(t),$$

and we recover the conclusions of the classical Maximum Principle of Pontryagin. (An extension is obtained when  $U(t)$  not bounded, see [10].) When  $f$  is not assumed to be differentiable, but merely locally Lipschitz with respect to  $x$ , there is a variant of the theorem in which the adjoint inclusion is expressed as follows:

$$-\dot{p}(t) \in \partial_C \langle p(t), f(t, \cdot, u_*(t)) \rangle (x_*(t)),$$

where the generalized gradient  $\partial_C$  (see Sect. “[Elements of Nonsmooth Analysis](#)”) is taken with respect to the  $x$  variable (note the connection to the standard adjoint equation above). This is an early form of the nonsmooth maximum principle [13].

## The Calculus of Variations

When we take  $f(t, x, u) = u$ , the problem reduces to the *problem of Bolza* in the calculus of variations. The first growth condition is trivially satisfied, and the second coincides with the generalized Tonelli–Morrey growth condition introduced in [11]; in this way we recover the state-of-the-art necessary conditions for the generalized problem of Bolza, which include as a special case the multiplier rule for problems with pointwise and/or isoperimetric constraints.

## Differential Inclusions

When we specialize further by taking  $F(t, \cdot)$  to be the indicator of the graph of a multifunction  $M(t, \cdot)$ , we obtain the principal necessary conditions for the *differential inclusion problem*. These in turn lead to necessary conditions for *generalized control systems* [11].

## Mixed Constraints

Consider again the optimal control problem in Mayer form (no integral term in the cost), but now in the presence of mixed state/control pointwise constraints of the form  $(x(t), u(t)) \in \Omega$  a.e. for a given closed set  $\Omega$ . Obtaining general necessary conditions for such problems is a well-known challenge in the subject; see [33, 34, 46]. We treat this case by taking  $F(t, \cdot) = I_\Omega(\cdot)$ . Then the second growth condition reduces to the following geometric assumption: for every  $(x, u) \in \Omega$ , for every  $(\zeta, \psi) \in N_\Omega^P(x, u)$ , one has

$$\frac{|\zeta|(1 + \|D_u f(t, x, u)\|)}{|\psi|} \leq c |f(t, x, u)| + d(t).$$

By taking a suitable representation for  $\Omega$  in terms of functional equalities and/or inequalities, sufficient conditions in terms of rank can be adduced which imply this property, leading to explicit multiplier rules (see [10] for details). With an appropriate ‘transversal intersection’ condition, we can also treat the case in which *both* the constraints  $(x(t), u(t)) \in \Omega$  and  $u(t) \in U(t)$  are present.

## Sensitivity

The multiplier functions  $p$  that appear in necessary conditions such as the ones above play a central role in analyzing the sensitivity of optimal control problems that depend on parameters. To take but one example, consider the presence of a perturbation  $\alpha(\cdot)$  in the dynamics of the problem:

$$\dot{x}(t) = f(t, x(t), u(t)) + \alpha(t) \text{ a.e.}$$

Clearly the minimum in the problem depends upon  $\alpha$ ; we denote it  $V(\alpha)$ . The function  $V$  is an example of what is referred to as a *value function*. Knowledge of the derivative of  $V$  would be highly relevant in studying the sensitivity of the problem to perturbations (errors) in the dynamics. Generally, however, value functions are not differentiable, so instead one uses nonsmooth analysis; the multipliers  $p$  give rise to estimates on the generalized gradient of  $V$ . We refer to [15, 16] for examples and references.

## Verification Functions

We consider now a special case of the problem considered in the preceding section: to minimize the integral cost functional

$$J(x, u) := \int_a^b F(t, x(t), u(t)) dt$$

subject to the prescribed boundary conditions

$$x(a) = A, \quad x(b) = B$$

and the standard control dynamics

$$\dot{x}(t) = f(t, x(t), u(t)) \text{ a.e., } u(t) \in U(t) \text{ a.e.}$$

Suppose now that a candidate  $(x_*, u_*)$  has been identified as a possible solution to this problem (perhaps through a partial analysis of the necessary conditions, or otherwise). An elementary yet powerful way (traceable to Legendre) to prove that  $(x_*, u_*)$  actually does solve the problem is to exhibit a function  $\phi(t, x)$  such that:

$$\begin{aligned} F(t, x, u) &\geq \phi_t(t, x) + \langle \phi_x(t, x), f(t, x, u) \rangle \\ \forall t \in [a, b], (x, u) &\in \mathbb{R}^n \times U(t), \end{aligned}$$

with equality along  $(t, x_*(t), u_*(t))$ .

The mere existence of such a function verifies that  $(x_*, u_*)$  is optimal, as we now show. For any admissible state/control pair  $(x, u)$ , we have

$$\begin{aligned} F(t, x(t), u(t)) &\geq \phi_t(t, x(t)) + \langle \phi_x(t, x(t)), \dot{x}(t) \rangle \\ &= \frac{d}{dt} \{ \phi(t, x(t)) \} \end{aligned}$$

$$\begin{aligned} \Rightarrow J(x, u) &= \int_a^b F(t, x(t), u(t)) dt \geq \phi(t, x(t)) \Big|_{t=a}^{t=b} \\ &= \phi(b, B) - \phi(a, A). \end{aligned}$$

But this lower bound on  $J$  holds with equality when  $(x, u) = (x_*, u_*)$ , which proves that  $(x_*, u_*)$  is optimal.

In this argument, we have implicitly supposed that the *verification function*  $\phi$  is smooth. It is a fact that if



we limit ourselves to smooth verification functions, then there may not exist such a  $\varphi$  (even when the problem itself has smooth data). However, if we admit nonsmooth (locally Lipschitz) verification functions, then (under mild hypotheses on the data) the existence of a verification function  $\varphi$  becomes necessary and sufficient for  $(x_*, u_*)$  to be optimal.

An appropriate way to extend the smooth inequality above uses the generalized gradient  $\partial_C \phi$  of  $\varphi$  as follows:

$$F(t, x, u) \geq \theta + \langle \zeta, f(t, x, u) \rangle \quad \forall (\theta, \zeta) \in \partial_C \phi(t, x),$$

together with the requirement

$$J(x_*, u_*) = \phi(b, B) - \phi(a, A).$$

It is a feature of this approach to proving optimality that it extends readily to more general problems, for example those involving unilateral state constraints, boundary costs, or isoperimetric conditions. It can also be interpreted in terms of duality theory, as shown notably by Vinter. We refer to [26] for details.

The basic inequality that lies at the heart of this approach is of Hamilton–Jacobi type, and the fact that we are led to consider nonsmooth verification functions is related to the phenomenon that Hamilton–Jacobi Equations may not admit smooth solutions. This is one of the main themes of the next section.

## Dynamic Programming and Viscosity Solutions

### The Minimal Time Problem

By a *trajectory* of the standard control system

$$\dot{x}(t) = f(x(t), u(t)) \text{ a.e., } u(t) \in U \text{ a.e.}$$

we mean a state function  $x(\cdot)$  corresponding to some choice of admissible control function  $u(\cdot)$ . The *minimal time problem* refers to finding a trajectory that reaches the origin as quickly as possible. Thus we seek the least  $T \geq 0$  admitting a control function  $u(\cdot)$  on  $[0, T]$  having the property that the resulting trajectory  $x$  satisfies  $x(T) = 0$ . We proceed now to describe the well-known *dynamic programming* approach to solving the problem.

We begin by introducing the *minimal time function*  $T(\cdot)$ , defined on  $\mathbb{R}^n$  as follows:  $T(\alpha)$  is the least time  $T \geq 0$  such that some trajectory  $x(\cdot)$  satisfies

$$x(0) = \alpha, x(T) = 0.$$

An issue of *controllability* arises here: Is it always possible to steer  $\alpha$  to 0 in finite time? When such is not

the case, then in accord with the usual convention we set  $T(\alpha) = +\infty$ .

The *principle of optimality* is the dual observation that if  $x(\cdot)$  is any trajectory, then we have, for  $s < t$ ,

$$T(x(t)) - T(x(s)) \geq s - t.$$

Equivalently, the function

$$t \mapsto T(x(t)) + t$$

is increasing. Furthermore, if  $x$  is optimal, then the same function is constant.

Let us explain this in other terms: if  $x(\cdot)$  is an optimal trajectory joining  $\alpha$  to 0, then

$$T(x(t)) = T(\alpha) - t \quad \text{for } 0 \leq t \leq T(\alpha),$$

since an optimal trajectory from the point  $x(t)$  is furnished by the truncation of  $x(\cdot)$  to the interval  $[t, T(\alpha)]$ . If  $x(\cdot)$  is any trajectory, then the inequality

$$T(x(t)) \geq T(\alpha) - t$$

is a reflection of the fact that in going to the point  $x(t)$  from  $\alpha$  (in time  $t$ ), we may have acted optimally (in which case equality holds) or not (then inequality holds).

Since  $t \mapsto T(x(t)) + t$  is increasing, we expect to have

$$\langle \nabla T(x(t)), \dot{x}(t) \rangle + 1 \geq 0,$$

with equality when  $x(\cdot)$  is an optimal trajectory. The possible values of  $\dot{x}(t)$  for a trajectory being precisely the elements of the set  $f(x(t), U)$ , we arrive at

$$\min_{u \in U} \langle \nabla T(x), f(x(t), u) \rangle + 1 = 0. \quad (4)$$

We define the (lower) *Hamiltonian function*  $h$  as follows:

$$h(x, p) := \min_{u \in U} \langle p, f(x, u) \rangle.$$

In terms of  $h$ , the partial differential equation obtained above reads

$$h(x, \nabla T(x)) + 1 = 0, \quad (5)$$

a special case of the *Hamilton–Jacobi Equation*.

Here is the first step in the dynamic programming heuristic: use the Hamilton–Jacobi Equation (5), together with the boundary condition  $T(0) = 0$ , to find  $T(\cdot)$ . How will this help us find the optimal trajectory?

To answer this question, we recall that an optimal trajectory is such that equality holds in (4). This suggests the

following procedure: for each  $x$ , let  $k(x)$  be a point in  $U$  satisfying

$$\min_{u \in U} \langle \nabla T(x), f(x, u) \rangle = \langle \nabla T(x), f(x, k(x)) \rangle = -1. \quad (6)$$

Then, if we construct  $x(\cdot)$  via the initial-value problem

$$\dot{x}(t) = f(x(t), k(x(t))), \quad x(0) = \alpha, \quad (7)$$

we will have a trajectory that is optimal (from  $\alpha$ )!

Here is why: Let  $x(\cdot)$  satisfy (7); then  $x(\cdot)$  is a trajectory, and

$$\begin{aligned} \frac{d}{dt} T(x(t)) &= \langle \nabla T(x(t)), \dot{x}(t) \rangle \\ &= \langle \nabla T(x(t)), f(x(t), k(x(t))) \rangle = -1. \end{aligned}$$

Integrating, we find

$$T(x(t)) = T(\alpha) - t,$$

which implies that at  $t = T(\alpha)$ , we must have  $x = 0$ . Therefore  $x(\cdot)$  is an optimal trajectory.

Let us stress the important point that  $k(\cdot)$  generates the optimal trajectory from *any* initial value  $\alpha$  (via (7)), and so constitutes what can be considered the ultimate solution for this problem: an *optimal feedback synthesis*. There can be no more satisfying answer to the problem: If you find yourself at  $x$ , just choose the control value  $k(x)$  to approach the origin as fast as possible. This goes well beyond finding a single open-loop optimal control.

Unfortunately, there are serious obstacles to following the route that we have just outlined, beginning with the fact that  $T$  is nondifferentiable, as simple examples show, even when it is finite everywhere (which it generally fails to be).

We will therefore have to examine anew the argument that led to the Hamilton–Jacobi Equation (5), which, in any case, will have to be recast in some way to accommodate nonsmooth solutions. Having done so, will the generalized Hamilton–Jacobi Equation admit  $T$  as the unique solution?

The next step (after characterizing  $T$ ) offers fresh difficulties of its own. Even if  $T$  were smooth, there would be in general no *continuous* function  $k(\cdot)$  satisfying (6) for each  $x$ . The meaning and existence of a trajectory  $x(\cdot)$  generated by  $k(\cdot)$  via the differential Eq. (7), in which the right-hand side is discontinuous in the state variable, is therefore problematic in itself.

The intrinsic difficulties of this approach to the minimal-time problem have made it a historical focal point of activity in differential equations and control, and it is only recently that fully satisfying answers to all the questions raised above have been found. We begin with generalized solutions of the Hamilton–Jacobi Equation.

## Subdifferentials and Viscosity Solutions

We shall say that  $\varphi$  is a *proximal solution* of the Hamilton–Jacobi Equation (5) provided that

$$h(x, \partial_P \phi(x)) = -1 \quad \forall x \in \mathbb{R}^n, \quad (8)$$

a ‘multivalued equation’ which means that for all  $x$ , for all  $\zeta \in \partial_P \phi(x)$  (if any), we have  $h(x, \zeta) = -1$ . (Recall that the proximal subdifferential  $\partial_P \phi$  was defined in Sect. “[Elements of Nonsmooth Analysis](#)”.)

Note that the equation holds automatically at a point  $x$  for which  $\partial_P \phi(x)$  is empty; such points play an important role, in fact, as we now illustrate. Consider the case in which  $f(x, U)$  is equal to the unit ball for all  $x$ , in dimension  $n = 1$ . Then  $h(x, p) \equiv -|p|$  (and the equation is of *eikonal* type). Let us examine the functions  $\phi_1(x) := -|x|$  and  $\phi_2(x) := |x|$ ; they both satisfy  $h(x, \nabla \phi(x)) = -1$  at all points  $x \neq 0$ , since (for each of these functions) the proximal subdifferential at points different from 0 reduces to the singleton consisting of the derivative. However, we have (see the exercise in the summary of nonsmooth analysis)

$$\partial_P \phi_1(0) = \emptyset, \quad \partial_P \phi_2(0) = [-1, 1],$$

and it follows that  $\phi_1$  is (but  $\phi_2$  is not) a proximal solution of the Hamilton–Jacobi Equation (8).

A lesson to be drawn from this example is that in defining generalized solutions we need to look closely at the differential behavior at specific and individual points; we cannot argue in an ‘almost everywhere’ fashion, or by ‘smearing’ via integration (as is done for linear partial differential equations via distributional derivatives).

Proximal solutions are just one of the ways to define generalized solutions of certain partial differential equations, a topic of considerable interest and activity, and one which seems to have begun with the Hamilton–Jacobi Equation in every case. The first ‘subdifferential type’ of definition was given by the author in the 1970s, using generalized gradients and for locally Lipschitz solutions. While no uniqueness theorem holds for that solution concept, it was shown that the value function of the associated optimal control problem is a solution (hence existence holds), and is indeed a special solution: it is the maximal one. In 1980 A. I. Subbotin defined his ‘minimax solutions’, which are couched in terms of Dini derivatives rather than subdifferentials, and which introduced the important feature of being ‘two-sided’. This work featured existence and uniqueness in the class of Lipschitz functions, the solution being characterized as the value of a differential game. Subsequently, M. Crandall and P.-L. Lions incorporated both subdifferentials and two-sidedness in their *viscosity solutions*, a theory which they developed for merely

continuous functions. In the current context, and under mild hypotheses on the data, it can be shown that minimax, viscosity, and proximal solutions all coincide [19,24].

Recall that our goal (within the dynamic programming approach) is to characterize the minimal time function. This is now attained, as shown by the following (we omit the hypotheses; see [63], and also the extensive discussion in Bardi and Capuzzo-Dolcetta [4]):

**Theorem 2** *There exists a unique lower semicontinuous function  $\phi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  bounded below on  $\mathbb{R}^n$  and satisfying the following:*

**[HJ equation]**  $h(x, \partial_P \phi(x)) = -1 \forall x \neq 0$ ;

**[Boundary condition]**  $\phi(0) = 0$  and  $h(0, \partial_P \phi(0)) \geq -1$ .

*That unique function is  $T(\cdot)$ .*

The proof of this theorem is based upon proximal characterizations of certain monotonicity properties of trajectories related to the inequality forms of the Hamilton–Jacobi Equation (see Sect. 4.7 of [24]). The fact that monotonicity is closely related to the solution of the minimal time problem is already evident in the following elementary assertion: a trajectory  $x$  joining  $\alpha$  to 0 is optimal if the rate of change of the function  $t \mapsto T(x(t))$  is  $-1$  a.e.

The characterization of  $T$  given by the theorem can be applied to verify the validity of a general conjecture regarding the time-optimal trajectories from arbitrary initial values. This works as follows: we use the conjecture to calculate the supposedly optimal time  $T(\alpha)$  from any initial value  $\alpha$ , and then we see whether or not the function  $T$  constructed in this way satisfies the conditions of the theorem. If so, then (by uniqueness)  $T$  is indeed the minimal time function and the conjecture is verified (the same reasoning as in the preceding section is in fact involved here). Otherwise, the conjecture is necessarily incorrect (but the way in which it fails can provide information on how it needs to be modified; this is another story).

Another way in which the reasoning above is exploited is to discretize the underlying problem as well as the Hamilton–Jacobi Equation for  $T$ . This gives rise to the backwards recursion numerical method developed and popularized by Bellman under the name of *dynamic programming*; of course, the approach applies to problems other than minimal time.

With respect to the goal of finding an optimal feedback synthesis for our problem, we have reached the following point in our quest: given that  $T$  satisfies the proximal Hamilton–Jacobi Equation  $h(x, \partial_P T(x)) = -1$ , which can be written in the form

$$\min_{u \in U} \langle \zeta, f(x(t), u) \rangle = -1 \quad \forall \zeta \in \partial_P T(x), \forall x \neq 0, \quad (9)$$

how does one proceed to construct a feedback  $k(x)$  having the property that any trajectory  $x$  generated by it via (7) is such that  $t \mapsto T(x(t))$  decreases at a unit rate? There will also arise the issue of defining the very sense of (7) when  $k(\cdot)$  is discontinuous, as it must be in general.

This is a rather complex question to answer, and it turns out to be a special case of the issue of designing stabilizing feedback (where, instead of  $T$ , we employ a *Lyapunov function*). We shall address this below, and return there to the minimal time synthesis, which will be revealed to be a special case of the procedure. We need first to examine the concept of Lyapunov function.

## Lyapunov Functions

In this section we consider the standard control system

$$\dot{x}(t) = f(x(t), u(t)) \text{ a.e.}, \quad u(t) \in U \text{ a.e.},$$

under a mild local Lipschitz hypothesis: for every bounded set  $S$  in  $\mathbb{R}^n$  there exists a constant  $K = K(S)$  such that

$$|f(x, u) - f(x', u)| \leq K|x - x'| \quad \forall x, x' \in S, \quad u \in U.$$

We also suppose that  $f(0, U)$  is bounded.

A point  $\alpha$  is *asymptotically guidable* to the origin if there is a trajectory  $x$  satisfying  $x(0) = \alpha$  and  $\lim_{t \rightarrow \infty} x(t) = 0$ . When every point has this property, and when additionally the origin has the familiar local stability property known as *Lyapunov stability*, it is said in the literature to be GAC: (open loop) *globally asymptotically controllable* (to 0). A well-known *sufficient* condition for this property is the existence of a smooth ( $C^1$ , say) pair of functions

$$V : \mathbb{R}^n \rightarrow \mathbb{R}, \quad W : \mathbb{R}^n \setminus \{0\} \rightarrow \mathbb{R}$$

satisfying the following conditions:

### 1. Positive Definiteness:

$$V(x) > 0 \text{ and } W(x) > 0 \quad \forall x \neq 0, \text{ and } V(0) \geq 0.$$

### 2. Properness: The sublevel sets $\{x : V(x) \leq c\}$ are bounded $\forall c$ .

### 3. Weak Infinitesimal Decrease:

$$\inf_{u \in U} \langle \nabla V(x), f(x, u) \rangle \leq -W(x) \quad x \neq 0.$$

The last condition asserts that  $V$  decreases in *some* available direction. We refer to  $V$  as a (weak) *Lyapunov function*; it is also referred to in the literature as a *control Lyapunov function*.

It is a fact, as demonstrated by simple examples (see [17] or [57]), that the existence of a smooth function  $V$  with the above properties fails to be a necessary condition for global asymptotic controllability; that is, the familiar converse Lyapunov theorems of Massera, Barbashin and Krasovskii, and Kurzweil (in the setting of a differential equation with no control) do not extend to this weak controllability setting, at least not in smooth terms. This may be a rather general phenomenon, in view of the following result [22], which holds under the additional hypothesis that the sets  $f(x, U)$  are closed and convex:

**Theorem 3** *If the system admits a  $C^1$  weak Lyapunov function, then it has the following surjectivity property: for every  $\epsilon > 0$ , there exists  $\delta > 0$  such that  $f(B(0, \epsilon), U) \supset B(0, \delta)$ .*

It is not difficult to check that the nonholonomic integrator system (see Sect. “Introduction”) is GAC. Since it fails to have the surjectivity property described in the theorem, it cannot admit a smooth Lyapunov function.

It is natural therefore to seek to weaken the smoothness requirement on  $V$  so as to obtain a necessary (and still sufficient) condition for a system to be GAC. This necessitates the use of some construct of nonsmooth analysis to replace the gradient of  $V$  that appears in the infinitesimal decrease condition. In this connection we use the proximal subgradient (Sect. “Elements of Nonsmooth Analysis”)  $\partial_P V(x)$ , which requires only that the (extended-valued) function  $V$  be lower semicontinuous. In proximal terms, the Weak Infinitesimal Decrease condition becomes

$$\sup_{\zeta \in \partial_P V(x)} \inf_{u \in U} \langle \zeta, f(x, u) \rangle \leq -W(x) \quad x \neq 0.$$

Note that this last condition is trivially satisfied when  $x$  is such that  $\partial_P V(x)$  is empty, in particular when  $V(x) = +\infty$ . (The supremum over the empty set is  $-\infty$ .) A *general Lyapunov pair*  $(V, W)$  refers to extended-valued lower semicontinuous functions  $V: \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  and  $W: \mathbb{R}^n \setminus \{0\} \rightarrow \mathbb{R} \cup \{+\infty\}$  satisfying the positive definiteness and properness conditions above, together with proximal weak infinitesimal decrease.

The following is proved in [24].

**Theorem 4** *Let  $(V, W)$  be a general Lyapunov pair for the system. Then any  $\alpha \in \text{dom } V$  is asymptotically guidable to 0.*

It follows from the theorem that the existence of a lower semicontinuous Lyapunov pair  $(V, W)$  with  $V$  everywhere finite-valued implies the global asymptotic guidability to 0 of the system. When  $V$  is not continuous, this does not

imply Lyapunov stability at the origin, however, so it cannot characterize global asymptotic controllability. An early and seminal result due to Sontag [56] considers *continuous* functions  $V$ , with the infinitesimal decrease condition expressed in terms of Dini derivatives. Here is a version of that result in proximal subdifferential terms:

**Theorem 5** *The system is GAC if and only if there exists a continuous Lyapunov pair  $(V, W)$ .*

There is an advantage to being able to replace continuity in such a result by a stronger regularity property (particularly in connection with using  $V$  to design a stabilizing feedback, as we shall see). Of course, we cannot assert smoothness, as pointed out above. In [20] it was shown that certain locally Lipschitz *value functions* give rise to *practical* Lyapunov functions, that is, assuring stable controllability to arbitrary neighborhoods of 0. Building upon this, L. Rifford [49] was able to combine a countable family of such functions in order to construct a global *locally Lipschitz* Lyapunov function. This answered a long-standing open question in the subject. Rifford [51] also went on to show the existence of a *semiconcave* Lyapunov function, a stronger property (familiar in pde’s) whose relevance to feedback construction will be seen in the next section.

The role of Lyapunov functions in characterizing various types of convergence to 0 (including guidability in finite time) is discussed in detail in [18].

## Stabilizing Feedback

We now address the issue of finding a *stabilizing feedback control*  $k(x)$ ; that is, a function  $k$  with values in  $U$  such that all solutions of the differential equation

$$\dot{x} = g(x), \quad x(0) = \alpha, \quad \text{where } g(x) := f(x, k(x)) \quad (10)$$

converge to 0 (for all values of  $\alpha$ ) in a suitable sense. Here, the origin is supposed to be an equilibrium of the system; to be precise, we take  $0 \in U$  and  $f(0, 0) = 0$ . When such a  $k$  exists, the system is termed *stabilizable*.

A necessary condition for the system to be stabilizable is that it be GAC. A central question in the subject has been whether this is sufficient as well. An early observation of Sontag and Sussmann [58] showed that the answer is negative if one requires the feedback  $k$  to be continuous (which provides the easiest classical way of interpreting the differential equation that appears in (10)). Later, Brockett showed that the nonholonomic integrator fails to admit a continuous stabilizing feedback.

One is therefore led to consider the use of discontinuous feedback, together with the attendant need to define an appropriate solution concept for a differential equation

in which the dynamics fail to be continuous in the state. The best-known solution concept in this regard is that of Filippov; it turns out, however, that the nonholonomic integrator fails to admit a (discontinuous) feedback which stabilizes it in the Filippov sense [32,55]. Clarke, Ledyaev, Sontag and Subbotin [21] gave a positive answer when the (discontinuous) feedbacks are implemented in the *closed-loop system sampling* sense (also referred to as *sample-and-hold*). We proceed now to describe the sample-and-hold implementation of a feedback.

Let  $\pi = \{t_i\}_{i \geq 0}$  be a partition of  $[0, \infty)$ , by which we mean a countable, strictly increasing sequence  $t_i$  with  $t_0 = 0$  such that  $t_i \rightarrow +\infty$  as  $i \rightarrow \infty$ . The *diameter* of  $\pi$ , denoted  $\text{diam}(\pi)$ , is defined as  $\sup_{i \geq 0} (t_{i+1} - t_i)$ . Given an initial condition  $x_0$ , the  $\pi$ -trajectory  $x(\cdot)$  corresponding to  $\pi$  and an arbitrary feedback law  $k: \mathbb{R}^n \rightarrow U$  is defined in a step-by-step fashion as follows. Between  $t_0$  and  $t_1$ ,  $x$  is a classical solution of the differential equation

$$\dot{x}(t) = f(x(t), k(x_0)), \quad x(0) = x_0, \quad t_0 \leq t \leq t_1.$$

(In the present context we have existence and uniqueness of  $x$ , and blow-up cannot occur.) We then set  $x_1 := x(t_1)$  and restart the system at  $t = t_1$  with control value  $k(x_1)$ :

$$\dot{x}(t) = f(x(t), k(x_1)), \quad x(t_1) = x_1, \quad t_1 \leq t \leq t_2,$$

and so on in this fashion. The trajectory  $x$  that results from this procedure is an actual state trajectory corresponding to a piecewise constant open-loop control; thus it is a physically meaningful one. When results are couched in terms of  $\pi$ -trajectories, the issue of defining a solution concept for discontinuous differential equations is effectively sidestepped. Making the diameter of the partition smaller corresponds to increasing the sampling rate in the implementation.

We remark that the use of possibly discontinuous feedback has arisen in other contexts. In linear time-optimal control, one can find discontinuous feedback syntheses as far back as the classical book of Pontryagin et al. [47]; in these cases the feedback is invariably piecewise constant relative to certain partitions of state space, and solutions either follow the switching surfaces or cross them transversally, so the issue of defining the solution in other than a classical sense does not arise. Somewhat related to this is the approach that defines a multivalued feedback law [6]. In stochastic control, discontinuous feedbacks are the norm, with the solution understood in terms of stochastic differential equations. In a similar vein, in the control of certain linear partial differential equations, discontinuous feedbacks can be interpreted in a distributional sense. These cases are all unrelated to the one under

discussion. We remark too that the use of discontinuous pursuit strategies in differential games [41] is well-known, together with examples to show that, in general, it is not possible to achieve the result of a discontinuous optimal strategy to within any tolerance by means of a continuous strategy (thus there can be a positive unbridgeable gap between the performance of continuous and discontinuous feedbacks).

It is natural to say that a feedback  $k(x)$  (continuous or not) *stabilizes* the system in the sample-and-hold sense provided that for every initial value  $x_0$ , for all  $\epsilon > 0$ , there exists  $\delta > 0$  and  $T > 0$  such that whenever the diameter of the partition  $\pi$  is less than  $\delta$ , then the corresponding  $\pi$ -trajectory  $x$  beginning at  $x_0$  satisfies

$$\|x(t)\| \leq \epsilon \quad \forall t \geq T.$$

The following theorem is proven in [21].

**Theorem 6** *The system is open loop globally asymptotically controllable if and only if there exists a (possibly discontinuous) feedback  $k: \mathbb{R}^n \rightarrow U$  which stabilizes it in the sample-and-hold sense.*

The proof of the theorem used the method of *proximal aiming*, which can be viewed as a geometric version of the Lyapunov technique. We now discuss how to define stabilizing feedbacks if one has in hand a sufficiently regular Lyapunov function.

### The Smooth Case

We begin with the case in which a  $C^1$  smooth Lyapunov function exists, where a very natural approach can be used. For  $x \neq 0$ , we simply define  $k(x)$  to be any element  $u \in U$  satisfying

$$\langle \nabla V(x), f(x, u) \rangle \leq -W(x)/2.$$

Note that at least one such  $u$  does exist, in light of the Infinitesimal Decrease Condition. It is then elementary to prove [18] that the pointwise feedback  $k$  described above stabilizes the system in the sample-and-hold sense.

We remark that Rifford [50] has shown that the existence of a smooth Lyapunov pair is equivalent to the existence of a locally Lipschitz one satisfying Infinitesimal Decrease in the sense of generalized gradients (that is, with  $\partial_P V$  replaced by  $\partial_C V$ ), and that this in turn is equivalent to the existence of a stabilizing feedback in the Filippov (as well as sample-and-hold) sense.

### Semiconcavity

We have seen that a smooth Lyapunov function generates a stabilizing feedback in a simple and natural way.



But since a smooth Lyapunov function does not necessarily exist, we still require a way to handle the general case. It turns out that the smooth and general cases can be treated in a unified fashion through the notion of *semiconcavity*, which is a certain regularity property (not implying smoothness). Rifford has proven that any GAC system admits a semiconcave Lyapunov function; we shall see that this property permits a natural extension of the pointwise definition of a stabilizing feedback that was used in the smooth case.

A function  $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$  is said to be (globally) semiconcave provided that for every ball  $B(0, r)$  there exists  $\gamma = \gamma(r) \geq 0$  such that the function  $x \mapsto \phi(x) - \gamma|x|^2$  is (finite and) concave on  $B(0, r)$ . (Hence  $\phi$  is locally the sum of a concave function and a quadratic one.) Observe that any function of class  $C^2$  is semiconcave; also, any semiconcave function is locally Lipschitz, since both concave functions and smooth functions have that property. (There is a local definition of semiconcavity that we omit for present purposes.)

Semiconcavity is an important regularity property in partial differential equations (see for example [9]). The fact that the semiconcavity of a Lyapunov function  $V$  turns out to be useful in stabilization is a new observation, and may be counterintuitive:  $V$  often has an interpretation in terms of energy, and it may seem more appropriate to seek a *convex* Lyapunov function  $V$ . We proceed now to explain why semiconcavity is a highly desirable property, and why a convex  $V$  would be of less interest (unless it were smooth, but then it would be semiconcave too).

Recall the ideal case discussed above, in which (for a smooth  $V$ ) we select a function  $k(x)$  such that

$$\langle \nabla V(x), f(x, k(x)) \rangle \leq -W(x)/2.$$

How might this appealing idea be adapted to the case in which  $V$  is nonsmooth? We cannot use the proximal subdifferential  $\partial_P V(x)$  directly, since it may be empty for some values of  $x$ . We are led to consider the limiting subdifferential  $\partial_L V(x)$  (see Sect. “[Elements of Nonsmooth Analysis](#)”), which is nonempty when  $V$  is locally Lipschitz. By passing to the limit, the Weak Infinitesimal Decrease Condition for proximal subgradients implies the following:

$$\inf_{u \in U} \langle f(x, u), \zeta \rangle \leq -W(x) \quad \forall \zeta \in \partial_L V(x), \quad \forall x \neq 0.$$

Accordingly, let us consider the following idea: for each  $x \neq 0$ , choose some element  $\zeta \in \partial_L V(x)$ , then choose  $k(x) \in U$  such that

$$\langle f(x, k(x)), \zeta \rangle \leq -W(x)/2.$$

Does this lead to a stabilizing feedback, when (of course) the discontinuous differential equation is interpreted in the sample-and-hold sense? When  $V$  is smooth, the answer is ‘yes’, as we have seen. But when  $V$  is merely locally Lipschitz, a certain ‘dithering’ phenomenon may arise to prevent  $k$  from being stabilizing. However, if  $V$  is semiconcave (locally on  $\mathbb{R}^n \setminus \{0\}$ ), this does not occur, and stabilization is guaranteed. This explains the interest in finding a semiconcave Lyapunov function.

When  $V$  is a locally Lipschitz Lyapunov function with no additional regularity (neither smooth nor semiconcave), then it can still be used for defining stabilizing feedback, but less directly. It is possible to *regularize*  $V$ : to approximate it by a semiconcave function through the process of *inf-convolution* (see [24]). This leads to *practical semiglobal* stabilizing feedbacks, that is, for any  $0 < r < R$ , we derive a feedback which stabilizes all initial values in  $B(0, R)$  to  $B(0, r)$  [20].

### Optimal Feedback

The strategy described above for defining stabilizing feedbacks via Lyapunov functions can also be applied to construct (nearly) optimal feedbacks as well as stabilizing ones. The key is to use an appropriate value function instead of an arbitrary Lyapunov function. We obtain in this way a unification of optimal control and feedback control, at least at the mathematical level, and as regards feedback design.

To illustrate, consider again the Minimal Time problem, at the point at which we had left it at the end of Sect. “[Lyapunov Functions](#)” (thus, unresolved as regards the time-optimal feedback synthesis). As pointed out there,  $T$  satisfies the Hamilton–Jacobi Equation: this yields Infinitesimal Decrease in subgradient terms. Consequently, if the Minimal Time function  $T$  happens to be finite everywhere as well as smooth or semiconcave, then we can use the same direct definition as above to design a feedback which (in the limiting sample-and-hold sense) produces trajectories along which  $T$  decreases at rate 1; that is, which are time-optimal. This of course yields the fastest possible stabilization (and in finite time). In general, however,  $T$  may lack such regularity, or (when the controllability to the origin is only asymptotic) not even be finite everywhere. Then it is necessary to apply an approximation (regularization) procedure in order to obtain a variant of  $T$ , and use that instead. When  $T$  is finite, we can obtain in this way an approximate time-optimal synthesis (to any given tolerance).

The whole approach described here can be carried out for a variety of optimal control contexts [26,45], and also

for finding optimal strategies in *differential games* [25]. It also carries over to problems in which *unilateral state constraints* are imposed:  $x(t) \in X$ , where  $X$  is a given closed set [26,28,29]. The issue of robustness, not discussed here, is particularly important in the presence of discontinuity; see [18,42,57].

### Future Directions

A lesson of the past appears to be that nonsmooth analysis is likely to be required whenever linearization is not adequate or is inapplicable. It seems likely therefore to accompany the subject of control theory as it sets out to conquer new nonlinear horizons, in ways that cannot be fully anticipated. Let us nonetheless identify a few directions for future work.

The extensions of most of the results cited above to problems on manifolds, or of tracking, or with partial information (as in adaptive control) remain to be carried out to a great extent. There are a number of currently evolving contexts not discussed above in which nonsmooth analysis is highly likely to play a role, notably *hybrid control*; an example here is provided by *multiprocesses* [8,31]. Distributed control (of pde's) is another area which requires development. There is also considerable work to be done on numerical implementation; in this connection see [36,37].

### Bibliography

- Artstein Z (1983) Stabilization with relaxed controls. *Nonlinear Anal TMA* 7:1163–1173
- Astolfi A (1996) Discontinuous control of nonholonomic systems. *Syst Control Lett* 27:37–45
- Astolfi A (1998) Discontinuous control of the Brockett integrator. *Eur J Control* 4:49–63
- Bardi M, Capuzzo-Dolcetta I (1997) Optimal control and viscosity solutions of Hamilton–Jacobi–Bellman equations. Birkhäuser, Boston
- Bardi M, Staicu V (1993) The Bellman equation for time-optimal control of noncontrollable nonlinear systems. *Acta Appl Math* 31:201–223
- Berkovitz LD (1989) Optimal feedback controls. *SIAM J Control Optim* 27:991–1006
- Borwein JM, Zhu QJ (1999) A survey of subdifferential calculus with applications. *Nonlinear Anal* 38:687–773
- Caines PE, Clarke FH, Liu X, Vinter RB (2006) A maximum principle for hybrid optimal control problems with pathwise state constraints. *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, 13–15 Dec 2006
- Cannarsa P, Sinestrari C (2004) Semiconcave Functions, Hamilton–Jacobi Equations, and Optimal Control. Birkhäuser, Boston
- Clarke F (2005) The maximum principle in optimal control. *J Cybern Control* 34:709–722
- Clarke F (2005) Necessary Conditions in Dynamic Optimization. *Mem Amer Math Soc* 173(816)
- Clarke FH (1973) Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations. Doctoral thesis, University of Washington
- Clarke FH (1976) The maximum principle under minimal hypotheses. *SIAM J Control Optim* 14:1078–1091
- Clarke FH (1983) Optimization and Nonsmooth Analysis. Wiley-Interscience, New York. Republished as: *Classics in Applied Mathematics*, vol 5. SIAM, 1990
- Clarke FH (1986) Perturbed optimal control problems. *IEEE Trans Aut Control* 31:535–542
- Clarke FH (1989) Methods of Dynamic and Nonsmooth Optimization. *Regional Conference Series in Applied Mathematics*, vol 57. SIAM, Philadelphia
- Clarke FH (2001) Nonsmooth analysis in control theory: a survey. *Eur J Control* 7:63–78
- Clarke FH (2004) Lyapunov functions and feedback in nonlinear control. In: de Queiroz MS, Malisoff M, Wolenski P (eds) *Optimal Control, Stabilization and Nonsmooth Analysis. Lecture Notes in Control and Information Sciences*, vol 301. Springer, New York, pp 267–282
- Clarke FH, Ledyaev YS (1994) Mean value inequalities in Hilbert space. *Trans Amer Math Soc* 344:307–324
- Clarke FH, Ledyaev YS, Rifford L, Stern RJ (2000) Feedback stabilization and Lyapunov functions. *SIAM J Control Optim* 39:25–48
- Clarke FH, Ledyaev YS, Sontag ED, Subbotin AI (1997) Asymptotic controllability implies feedback stabilization. *IEEE Trans Aut Control* 42:1394–1407
- Clarke FH, Ledyaev YS, Stern RJ (1998) Asymptotic stability and smooth Lyapunov functions. *J Differ Equ* 149:69–114
- Clarke FH, Ledyaev YS, Stern RJ, Wolenski PR (1995) Qualitative properties of trajectories of control systems: a survey. *J Dyn Control Syst* 1:1–48
- Clarke FH, Ledyaev YS, Stern RJ, Wolenski PR (1998) *Nonsmooth Analysis and Control Theory. Graduate Texts in Mathematics*, vol 178. Springer, New York
- Clarke FH, Ledyaev YS, Subbotin AI (1997) The synthesis of universal pursuit strategies in differential games. *SIAM J Control Optim* 35:552–561
- Clarke FH, Nour C (2005) Nonconvex duality in optimal control. *SIAM J Control Optim* 43:2036–2048
- Clarke FH, Rifford L, Stern RJ (2002) Feedback in state constrained optimal control. *ESAIM Control Optim Calc Var* 7:97–133
- Clarke FH, Stern RJ (2005) Hamilton–Jacobi characterization of the state-constrained value. *Nonlinear Anal* 61:725–734
- Clarke FH, Stern RJ (2005) Lyapunov and feedback characterizations of state constrained controllability and stabilization. *Syst Control Lett* 54:747–752
- Clarke FH, Vinter RB (1984) On the conditions under which the Euler equation or the maximum principle hold. *Appl Math Optim* 12:73–79
- Clarke FH, Vinter RB (1989) Applications of optimal multiprocesses. *SIAM J Control Optim* 27:1048–1071
- Coron J-M, Rosier L (1994) A relation between continuous time-varying and discontinuous feedback stabilization. *J Math Syst Estim Control* 4:67–84
- de Pinho MR (2003) Mixed constrained control problems. *J Math Anal Appl* 278:293–307

34. Dmitruk AV (1993) Maximum principle for a general optimal control problem with state and regular mixed constraints. *Comp Math Model* 4:364–377
35. Ferreira MMA (2006) On the regularity of optimal controls for a class of problems with state constraints. *Int J Syst Sci* 37:495–502
36. Fontes FACC (2001) A general framework to design stabilizing nonlinear model predictive controllers. *Syst Control Lett* 42:127–143
37. Fontes FACC, Magni L (2003) Min-max model predictive control of nonlinear systems using discontinuous feedbacks. *New directions on nonlinear control. IEEE Trans Autom Control* 48:1750–1755
38. Hamzi B, Praly L (2001) Ignored input dynamics and a new characterization of control Lyapunov functions. *Autom J IFAC* 37:831–841
39. Ioffe AD, Tikhomirov V (1974) *Theory of Extremal Problems*. Nauka, Moscow. English translation: North-Holland, Amsterdam, 1979
40. Kellett CM, Teel AR (2005) On the robustness of  $\mathcal{KL}$ -stability for difference inclusions: smooth discrete-time Lyapunov functions. *SIAM J Control Optim* 44:777–800
41. Krasovskii NN, Subbotin AI (1988) *Game-Theoretical Control Problems*. Springer, New York
42. Ledyaeu YS, Sontag ED (1999) A Lyapunov characterization of robust stabilization. *Nonlinear Anal* 37:813–840
43. Milyutin AA, Osmolovskii NP (1998) *Calculus of Variations and Optimal Control*. Amer Math Soc, Providence
44. Neustadt LW (1976) *Optimization*. Princeton University Press, Princeton
45. Nobakhtian S, Stern RJ (2000) Universal near-optimal feedbacks. *J Optim Theory Appl* 107:89–122
46. Páles Z, Zeidan V (2003) Optimal control problems with set-valued control and state constraints. *SIAM J Optim* 14:334–358
47. Pontryagin LS, Boltyanskii RV, Gamkrelidze RV, Mischenko EF (1962) *The Mathematical Theory of Optimal Processes*. Wiley-Interscience, New York
48. Prieur C, Trélat E (2005) Robust optimal stabilization of the Brockett integrator via a hybrid feedback. *Math Control Signal Syst* 17:201–216
49. Rifford L (2000) Existence of Lipschitz and semiconcave control-Lyapunov functions. *SIAM J Control Optim* 39:1043–1064
50. Rifford L (2001) On the existence of nonsmooth control-Lyapunov functions in the sense of generalized gradients. *ESAIM Control Optim Calc Var* 6:539–611
51. Rifford L (2002) Semiconcave control-Lyapunov functions and stabilizing feedbacks. *SIAM J Control Optim* 41:659–681
52. Rifford L (2003) Singularities of viscosity solutions and the stabilization problem in the plane. *Indiana Univ Math J* 52:1373–1396
53. Rockafellar RT, Wets R (1998) *Variational Analysis*. Springer, New York
54. Rodríguez H, Astolfi A, Ortega R (2006) On the construction of static stabilizers and static output trackers for dynamically linearizable systems, related results and applications. *Int J Control* 79:1523–1537
55. Ryan EP (1994) On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. *SIAM J Control Optim* 32:1597–1604
56. Sontag ED (1983) A Lyapunov-like characterization of asymptotic controllability. *SIAM J Control Optim* 21:462–471
57. Sontag ED (1999) Stability and stabilization: discontinuities and the effect of disturbances. In: Clarke FH, Stern RJ (eds) *Nonlinear Analysis, Differential Equations and Control*, NATO ASI Montreal 1998. Kluwer, Dordrecht, pp 551–598
58. Sontag ED, Sussmann HJ (1980) Remarks on continuous feedback. In: *Proc IEEE Conf Decis and Control Albuquerque*. IEEE Publications, Piscataway, pp 916–921
59. Subbotin AI (1995) *Generalized Solutions of First-Order PDEs*. Birkhäuser, Boston
60. Trélat E (2006) Singular trajectories and subanalyticity in optimal control and Hamilton–Jacobi theory. *Rend Semin Mat Univ Politec Torino* 64:97–109
61. Vinter RB (2000) *Optimal Control*. Birkhäuser, Boston
62. Warga J (1972) *Optimal Control of Differential and Functional Equations*. Academic Press, New York
63. Wolenski PR, Zhuang Y (1998) Proximal analysis and the minimal time function. *SIAM J Control Optim* 36:1048–1072

---

## Non-standard Analysis, An Invitation to

WEI-ZHE YANG

Department of Mathematics, National Taiwan University,  
Taipei, Taiwan

### Article Outline

[Glossary](#)  
[Definition of the Subject](#)  
[Introduction](#)  
[Pre-Robinson Infinitesimals](#)  
[Hyperreals](#)  
[General Idea of NSA](#)  
[Loeb Construction](#)  
[A Future for Non-standard Analysis?](#)  
[Bibliography](#)

### Glossary

**Hyperreal** A proper extension of the real numbers, containing infinities and infinitesimal numbers, such that a transfer principle allows first-order theorems in the reals to be extended therein.  
**Infinity** An element in an ordered field which is larger than  $1 + 1 + 1 + \dots + 1$  for any number of 1's.  
**Infinitesimal** An element in an ordered field whose absolute value is less than  $1/n$  for any positive integer  $n$ .  
**Non-archimedean** Infinitesimal or infinite values exist in an ordered field.  
**Transfer principle** A rule which transforms assertions about standard sets, mappings, etc., into one about in-

ternal sets, mappings. Intuitively, it is an elementary embedding between structures.

### Definition of the Subject

Generically speaking, non-standard analysis is any form of mathematics that relies on non-standard models and the transfer principle. However, it is likely safe to say that of the most common meaning for the term “non-standard analysis” is a rigorous formulation of analysis using infinitesimals, as an alternative to the analysis we all know in the  $\epsilon$ - $\delta$  formulation. Abraham Robinson is usually regarded as the first person to do this coherently [12].

It is sometimes claimed that non-standard analysis can achieve more than standard models [2]. This may not be entirely true, but philosophically, non-standard analysis justifies the standard pedagogical process – infinitesimals in Leibniz notation – in which most people learned calculus, and represents the completion of the a key developmental stage of science.

### Introduction

**Warning 1** *What an (general) undergraduate to a student of mathematics is what a (general) mathematician to a logician.*

This article, written by a (general) mathematician, is aimed at a reader with some mathematical maturity. There are different constructions of non-standard analysis, and we have chosen to start with the “ultrapower” construction of the semantic kind, which is not as difficult to comprehend as the first presentation.

### Differential = Infinitesimal Difference?

Suppose a freshman is questioned on the derivative  $\frac{dy}{dx}$  for the cubic function  $y = x^3$ . Then after the difference-quotient is found,

$$\frac{\Delta y}{\Delta x} = 3x^2 + 3x(\Delta x) + (\Delta x)^2;$$

he just says “substituting by  $\Delta x = 0$ ”,  $\frac{dy}{dx} = 3x^2$ . And the instructors often have trouble explaining the difference between “ $\Delta x$  set to 0” and “ $\lim_{\Delta x \rightarrow 0}$ ”.

It is estimated that, among the mathematical-literates (i. e., those who learnt calculus), less than 10% are competent with “ $\epsilon$ - $\delta$ -logy”. But, of course, they are assured that what they learnt is no nonsense. The calculus we learn is from Leibniz, not Seki Kowa (even if the textbook is in Japanese), and not even Newton, who must have the priority. And the discoveries of calculus and of rigorous analysis (chiefly of Cauchy), are some 150 years apart.

In essence, the success of Leibniz is due to his elegant and mnemonic way of notations, centered around the notion of differential. When a freshman understands that “the differential of  $y$ ”, written as  $dy$ , is “the infinitesimal difference of  $y$ ”, he deals with calculus problems just like doing routine high school algebra. Indeed calculus as envisaged by Leibniz, is a computation scheme in a number system, denoted by  ${}^*\mathbb{R}$  below. This system of “hyperreals” must include, besides the usual real numbers, infinitesimals. Leibniz conceived of the notion of infinitesimals out of necessity, because he was pre-Cauchy. He needed the infinitesimals (as differentials) in the calculus.

But from the very beginning many people, notably Bishop G. Berkeley, “discover impossibilities and contradictions”. Infinitesimals cannot be, simultaneously, zero and non-zero. If they are zero, the division is illegal; if they are not, when  $y = x^3$ ,  $\frac{dy}{dx} = 3x^2$  is wrong because  $3xdx + (dx)^2$  does not just evaporate away. Of course this is a contradiction that Leibniz, who was a first-rate philosopher-logician, could not fail to discover. Thus the calculus as a science was, for many people before Cauchy, is only recognized de facto and not de jure, just like the nations of Taiwan and Kosovo.

### Structure of This Article

In the next Sect. “Pre-Robinson Infinitesimals”, we will tell you how the people tried to handle infinitesimals before [12]. In Sect. “Hyperreals”, we define the hyperreal field using the ultrapower construction. Section “General Idea of NSA” builds a small framework of non-standard analysis using hyperreals, and Sect. “Loeb Construction” introduces the Loeb construction for Wiener processes in NSA.

### Pre-Robinson Infinitesimals

Herein we describe the work prior to Robinson (despite say the work of [13]), who is credited with being the creator of the field.

### Recap: Naive Set Notations

We assume the concept of a set without recourse to the axioms of set theory.

**Definition 1 (Binary Boolean Set-Operations)** For two sets  $A$  and  $B$ , there are these Boolean combinations of them:

$$A \cup B := \{x : x \in A \text{ or } x \in B\},$$

$$A \cap B := \{x : x \in A \text{ and } x \in B\},$$

$$A \setminus B := \{x : x \in A \text{ and } x \notin B\}.$$

## Non-standard Analysis, An Invitation to, Table 1

## Notations of this Chapter

Notation	Meaning
iff	if and only if
$\neg, \wedge, \vee, \Rightarrow$	"negation of ", "and ", "or ", "imply ".
$\forall, \exists$	"for all ", "exists "
$:=$	"defined to be equal to "
$\mathbb{R}, \mathbb{C}, \mathbb{Z}$	set of reals, complex numbers, integers
$\mathbb{N}$	set of natural numbers (zero excluded)
$(a \dots b), [a \dots b]$	open and closed intervals
$(a \dots b], [a \dots b)$	semi-open intervals
$\{ \quad \}$	the braces holding a set
$( \quad )$	angles holding a sequence
$\cup, \sqcup, \cap, \setminus$	set union, disjoint union, intersection, set difference
$\mathcal{P}(A)$	power set (set of all subsets including $A$ and $\emptyset$ )
$\mathbb{R}_{\beta}^{\mathbb{N}}$	set of all bounded sequences
$\mathbb{R}_{\mathcal{C}}^{\mathbb{N}}$	set of all convergent sequences
$\mathbb{R}_0^{\mathbb{N}}$	set of all sequences converging to zero(=vanishing)
$\mathbb{R}_{\mathcal{K}}^{\mathbb{N}}$	set of all eventually -zero(=nullifying) sequences
$\mathbb{R}^{\times}$	$= \mathbb{R} \setminus \{0\}$ , non-zero reals
$\mathbb{R}_+$	$[0 \dots \infty)$ , non -negative real number
$\mathbb{R}$	$= \mathbb{R} \sqcup \{-\infty, +\infty\}$ , $\mathbb{R}$ with signed infinities
$\mathbb{R}$	$= \mathbb{R} \sqcup \{\infty\}$ , $\mathbb{R}$ with unsigned infinity
$^*\mathbb{R}$	(one possible) hyperreal number system
$\mathcal{U}$	ultra -filter, also and denote (as superscript) "ultrapower construction"
$\sigma$	superscript denotes the set of standard elements
$\pi_{\mathcal{U}}$	denotes the projection for the ultrapower construction $\mathcal{U}$

**Definition 2 (Function)** Let  $A$  and  $B$  be two non-empty sets. If for any  $x \in A$ , there is correspondingly assigned one  $y \in B$ , and written as  $f(x)$ , then we call this assignment  $f$  a function. Its domain is  $A$ , codomain  $B$ . The notation is  $f: A \mapsto B$ , or  $f: x \in A \mapsto f(x) \in B$ . Also, for  $C \subset A$ , or for  $D \subset B$ , we write

$$f[C] := \{f(x) : x \in C\}.$$

$$f^{-1}[D] := \{x \in A : f(x) \in D\}.$$

If  $f[A] = B$ , then the function  $f$  is surjective. If  $f(x) = f(u)$  implies  $x = u$ , then the function  $f$  is injective. when the function is both injective and surjective, it is called bijective. Sometimes the codomain of a function is just understood, but suppressed, and the function is then written as an "indexed family"  $\{f_j : j \in I\}$ . In this case, we write  $f_j$  instead of  $f(j)$ , and  $I$ , the domain of definition, here  $I$ , is called the indexing-set, and is often omitted when no confusion is likely. When the indexing set is  $\mathbb{N}$  (and in a few other situations) the indexed family is called a **sequence**.

**Definition 3 (Cartesian-Product, Couple, Triple, and Sequence)** For two non-empty sets  $A, B$ , their Cartesian product is defined as

$$A \times B := \{\langle a, b \rangle : a \in A, \text{ and } b \in B\}.$$

Here "a couple", or "an ordered pair", is denoted by  $\langle a, b \rangle$ . An ordered  $n$ -tuple is analogously denoted like  $\langle x_1, x_2, \dots, x_n \rangle$ .

**Definition 4 (General Operations among a Family of Sets)** If we have an indexed family of sets  $(X_j : j \in I)$ , then we define:

$$\bigcup_{j \in I} X_j := \{x : x \in X_j, \text{ for some } j \in I\};$$

$$\bigcap_{j \in I} X_j := \{x : x \in X_j, \text{ for each } j \in I\};$$

$$\prod_{j \in I} X_j := \{g : g \text{ is a function, with } g(j) \in X_j, \text{ for all } j \in I\};$$

**Definition 5 (Quotient-Set and Equivalence Relation)** This is a review of the most difficult notion in naive set theory. In studying a non-empty set  $A$ , there often occur an



“equivalence relation”, denoted by, say, “ $\sim$ ”. This means we write “ $x \sim y$ ” to mean that these two elements  $x, y$  of  $A$  are “of the same kind”, “similar”, or “equivalent”, referring to something of our concern. If we collect all similar members into “bunches”, each bunch is called an “equivalence class” (of this relation  $\sim$ ). The set of (bunches=) equivalence-classes is called the quotient set of  $A$  with respect to the relation  $\sim$ , and written as  $\frac{A}{\sim}$ . This process of forming the quotient set  $\frac{A}{\sim}$  from the set  $A$ , is usually described as “identifying equivalent members”. Formally speaking: a binary relation  $\sim$  on  $A$  is called an equivalence relation iff it has the following three properties:

$r^\circ$  (Reflexivity): for any  $x \in A$ ,  $x \sim x$ .

$s^\circ$  (Symmetry): when  $\{x, y\} \subset A$ , and  $x \sim y$ , then  $y \sim x$ .

$t^\circ$  (Transitivity): when  $\{x, y, z\} \subset A$ , and  $x \sim y$ , and also  $y \sim z$ , then  $x \sim z$ .

For  $x \in A$ , its equivalence class (modulo  $\sim$ ), is then

$$\pi_{\sim}(x) := \{y \in A : x \sim y\}.$$

(It is a non-empty subset of  $A$ ). The quotient set is then:

$$\frac{A}{\sim} := \{\pi_{\sim}(x) : x \in A\}.$$

### Ordered Field Extension

The relation between the system  $\mathbb{R}$  of reals and the system  $^*\mathbb{R}$  of hyper-reals may be announced thus: First of all,  $^*\mathbb{R}$  should be a proper extension of the ordered-field  $\mathbb{R}$ . Secondly, one should also be able to extend, from  $\mathbb{R}$  to  $^*\mathbb{R}$ , the usual functions  $\exp$ ,  $\cos$ ,  $\sin$ , etc. This is one of the topics of this article, and we explain the meaning of “extension” in this subsection.

**Definition 6 (Commutative-Semigroup)** Let us consider the arithmetic operation  $+$  in  $\mathbb{R}$ . Though it is a two-variable function, just like, e. g.,  $f: f(x, y) = \sin(x - y) * (x^3 - y^2)$ , it is written “infix”. (So that  $x + y$  is “really”  $+(x, y)$ ). There are the noteworthy laws of associativity, and of commutativity:

$$\begin{aligned} \forall x, \forall y, \forall z, \\ (x + y) + z = (x + (y + z)); \quad x + y = y + x. \end{aligned}$$

(We mathematics student read  $\forall$  as “for all”). By abstraction, we call a binary operation (= two variable function) on a non-empty set  $A$ , a “commutative semigroup structure” on this set. This system, the set  $A$  together with the commutative semigroup structure, is called a commutative semigroup. We realize that: in  $\mathbb{R}$ , multiplication is another commutative semigroup structure.

**Definition 7 (Commutative-Group)** Now in  $\mathbb{R}$ , 0 is distinguished by the “unitality law”:

$$\forall x, x + 0 = 0 + x = x.$$

There is then a negation function, mapping  $x$  into its negative  $-x$ , and the reciprocal law holds:

$$x + (-x) = 0; \quad (-x) + x = 0.$$

By abstraction, we call a triple of a two variable function, a distinguished element, and a one-variable function, all on a non-empty set  $A$ , satisfying all the laws mentioned above, as a “commutative group structure”. And the set, together with this commutative group structure, is called a commutative group. We realize that:  $(\mathbb{R}, (+, 0, -))$  is a commutative group. As to the multiplication, the distinguished unity is 1, but the reciprocation fails at precisely 0. If we all the functions are restricted to the zero-deleted set

$$\mathbb{R}^\times := \mathbb{R} \setminus \{0\},$$

then we do have a commutative group structure on it.

**Definition 8 (Field)** Suppose  $F$  is a set of more than one elements. Let the triple  $(+, 0, -)$  be a commutative group structure, and  $(*, 1, ^{-1})$  be a commutative group structure on  $F^\times = F \setminus \{0\}$ . And suppose the distributive law holds:

$$(\forall x, \forall y, \forall z), \quad x * (y + z) = x * y + x * z.$$

Then this couple of triple is called a “field structure” on  $F$ . (Of course, there are, altogether, six items in the structure: two distinguished elements (=“null-variable functions”), two (single variable) functions, two two-variable functions). And the set  $F$ , together with this field structure, is called a field.

**Definition 9 (Total-Ordering, and Ordered Field)** A total-ordering on a set  $T$  is a binary relation on  $T$ , which satisfies the transitive law and the trichotomy law. For example, in  $\mathbb{R}$ , the relation “less than” is a total-ordering: the transitive law says that: if  $a < b$ , and  $b < c$ , then  $a < c$ , while the trichotomy law requires that: for any two members  $a, b$ , exactly one sentence among the following three is true:  $a < b$ ,  $a = b$ ,  $b < a$ . An ordered-field structure on a set  $K$  is a combination of a field structure, and an total-ordering  $<$ , with the additional requirement of compatibility law:

$$((0 < x) \wedge (0 < y)) \Rightarrow ((0 < x + y) \wedge (0 < x * y)).$$

$\wedge$  is read “and”, and  $\Rightarrow$  is read “imply”.  $K$  with this structure, is then an ordered-field. Obviously the set of all rational numbers  $\mathbb{Q}$ , together with the natural structure, (which

is called “inherited from  $\mathbb{R}$ ”), is also an ordered-field in this sense.

**Definition 10 (Mathematical System, Inheritance and Extension)** Returning to the instance of the set  $\mathbb{R}$ . We may consider the following substructure of five items: the three items of addition commutative-group structure, and the multiplication unital commutative semi-group structure, keeping in mind the distribution law binding them. This is a unital commutative-ring structure. A set  $K$  together with a unital commutative-ring structure on it is called a unital commutative-ring. And as is obvious from above discussion, by forgetting the two items in any field structure, we have a unital commutative-ring structure. So the latter is weakened from the former, while the former is strengthened from the latter. Essentially any one particular kind of structure is called a “category”. So within a category one system may be the extension of the other, or the other way round, the latter is a sub-system of the former. For example, the rational ordered field  $\mathbb{Q}$  is a sub-system of  $\mathbb{R}$ , and the real ordered field  $\mathbb{R}$  is an extension of  $\mathbb{Q}$ . This extension is proper, because  $\mathbb{R}$  is strictly bigger: there are irrational numbers. Now for example, the set  $\mathbb{Z}$  of all integers is also a unital commutative-ring, when its structure is inherited (with the obvious meaning, or technically “restricted”) from  $\mathbb{Q}$ . That is,  $\mathbb{Z}$  is a sub-unital commutative-ring of  $\mathbb{Q}$ . But it is not a subsystem of  $\mathbb{Q}$  as a field.

**Lemma 1 (Quotient Algebra)** *Let us mangle usual terminology a little and abbreviate “unital commutative-ring” as “algebra”, just for convenience in this discussion. If  $A$  is an algebra, then a subset  $B$  is called an ideal, if the following condition is met: for arbitrary  $x \in A$ ,  $y \in B$ ,  $z \in B$ ,*

$$y+z \in B, \quad x*y \in B, \quad (\text{which is same as}) \quad y*x \in B.$$

*Then we define a (modulo  $B$ ) relation  $\sim_B$  in  $A$ , by:*

$$(x \sim_B y) \quad \text{iff} \quad (x - y) \in B.$$

*And this is an equivalence relation. The equivalence-class of  $x \in A$  is then:*

$$x + B := \{x + y : y \in B\}.$$

*The quotient set  $\frac{A}{\sim_B}$  will be written as  $A \bmod B$ . In it naturally we define the operations by:*

$$(x + B) + (y + B) := (x + y) + B,$$

$$-(x + B) := (-x) + B,$$

$$(x + B) * (y + B) := (x * y) + B.$$

*(And endow it with the natural zero  $0 + B$ , the unit  $1 + B$ ). The simple result is: these indeed form an algebra structure.*

**Remark 1 (Other Category)** A one variable function is a particular kind of binary relation. And a two-variable function is a particular kind of ternary relation. Indeed a null-variable function (=a distinguished element), is a particular kind of unary relation. Usually structures consisting of these kind of functions, (bound by algebraic laws), are called “algebraic”. In topology, we study the topological structure, and in measure theory, we need the measurable structure, before strengthening it to the measure structure. And the set  $\mathbb{R}$  has on it a natural topological structure and a natural measure structure.

### Classical Augmented Infinities

**Definition 11 (Leibniz Program)** Extend the system  $\mathbb{R}$  of reals to the system  ${}^*\mathbb{R}$  of hyperreals, so as to contain infinitesimals and also infinities.

This question should be considered in the perspective of “structure of a system” explained in the last subsection. Indeed the NSA of Abraham Robinson is so powerful as to consider the super-structure which is all-encompassing. As far as the size or magnitude (=absolute-value) is concerned, hyperreals should be classified in three categories: infinitesimal, ordinary, and infinite. The reciprocal of an infinitesimal (assumed non-zero), must be infinite (=unlimited), and vice versa. While ordinary hyperreals are neither infinitesimal nor infinite. They are medium-sized, together with their reciprocal. The infinite hyperreals and the infinitesimals (zero excepted), are all non-standard real numbers. The non-zero standard real numbers are all included in the medium-sized hyperreals. Now in calculus, and especially at the beginning pages of complex variables, we sometimes talk about infinities. It turns out that they are standard infinities, and does not belong to the Leibnizian  ${}^*\mathbb{R}$  of hyperreals. Let us first try to add one or two points to  $\mathbb{R}$ , disregarding the infinitesimals for the present. Actually there are precisely three infinities, of two kinds: two signed infinities, and one unsigned infinity.

**Definition 12 (Signed Infinities)** We denote by  $+\infty$  the “positive infinity”, and by  $-\infty$  the “negative infinity”. They are added to the totally ordered set  $\mathbb{R}$ , so as to be, respectively, the maximum and the minimum of this order-extension of  $\mathbb{R}$ .

$$\mathbb{R} = \mathbb{R} \sqcup \{-\infty, +\infty\}.$$

(We use  $\sqcup$  here to emphasize that it is a disjoint-union.) So there is an order-isomorphism:

$$x \in \mathbb{R} \mapsto \frac{2}{\pi} \arctan(x) \in [-1 \dots 1], \quad (\text{a closed interval})$$

which is also an homeomorphism (=topological isomorphism). And this compact space  $\mathbb{R}$  is therefore called the two-point compactification of  $\mathbb{R}$ .

**Definition 13 (Supremum, Infimum)** In  $\mathbb{R}$ , any (non-empty) subset  $A$  has a supremum, and also an infimum, denoted respectively as  $\sup A$ ,  $\inf A$ .

We recall that “ $b$  is an upper-bound of  $A$ ” means that  $\forall x \in A$ ,  $x \leq b$ . And  $b = \sup A$  means:  $b$  is the least among all upper-bounds of  $A$ . In measure theory, most of the values are in the positive half of  $\mathbb{R}$ , that is, the set  $[0 \cdots +\infty] = \mathbb{R}_+$ . Finally there is one unsigned infinity, which is the only infinity of the one-point compactification

$$\mathbb{R} = \mathbb{R} \sqcup \{\infty\},$$

which is obtained by merging the two signed infinities  $\pm\infty$  of  $\mathbb{R}$  into one point. In  $\mathbb{R}$ , we have, for example:

$$-\infty < -921, -921 < -228, -228 < +\infty;$$

If the ordering were compatible with the merging, then, we would get

$$\infty < -921, -921 < -228, -228 < \infty;$$

and the transitive law is violated! This excludes the possibility of ordering the set  $\mathbb{R}$ . The space  $\mathbb{R}$  is homeomorphic to the unit circle, e.g. by the (one-dimensional!) stereographic projection

$$x \in \mathbb{R} \mapsto (\cos \theta, \sin \theta), \quad \text{where} \quad \theta = 2 * \arctan x - \frac{2}{\pi}.$$

Note: Here we use an unusual notation to distinguish  $\infty \in \mathbb{R}$  from  $+\infty \in \mathbb{R}$ . Actually these two infinities, one unsigned, the other signed, are never coexistent, and usually they both are denoted by the same simple  $\infty$ . The deficiency of these systems is that: arithmetic operations involving the infinities are only partially defined in the system, so that there are many undefined operations, so-called “undetermined form”:

$$(+\infty) + (-\infty); \quad (\pm\infty) * 0; \quad (\pm\infty) \div (\pm\infty);$$

Because of these no-definitions, the associative law is greatly compromised. So the use of these infinities is always just symbolic.

**Note 1** When  $\arctan(x)$  is defined for  $x \in \mathbb{R}$ , it is valued in  $[-\frac{\pi}{2} \dots \frac{\pi}{2}]$ , while for  $x \in \mathbb{R}$ , it is valued in  $\mathbb{R} \bmod (\pi)$ . Also  $\lim_{x \rightarrow 0} (7 - 2x)/x^4 = +\infty \in \mathbb{R}$ , while  $\lim_{x \rightarrow 0} (7 - 2x)/x^3 = \infty \in \mathbb{R}$ . Such distinctions are sometimes useful or necessary.

## Non-Archimedean Infinitesimal

Physically speaking, positive real numbers are used for measurements. The possibility of measurement is based on this simple fact: any length  $\lambda$  however long, may be measured against any length  $\epsilon$  however short. If, for example,  $\lambda = 1$  m, and  $\epsilon = 3.4$  mm, then  $0 < \lambda - n * \epsilon < \epsilon$ , with  $n = \text{floor}(\frac{\lambda}{\epsilon}) = 294$ , so that  $\lambda$  is a little bit more than  $n$  units of  $\epsilon$ . This basic fact is the

**Axiom 1 (of Archimedes)** For any  $\lambda \in \mathbb{R}_+$ , and any  $\epsilon \in \mathbb{R}_+$ , there is an  $n \in \mathbb{N}$ , such that  $n * \epsilon > \lambda$ .

It is precisely this axiom which denies the existence of the infinitesimals: a positive infinitesimal would not be usable as a unit for measurement.

**Definition 14 (Infinitesimal)** In an ordered field  $F$ , an element  $\epsilon$  is called infinitesimal, if for any  $n \in \mathbb{N}$ ,  $n * |\epsilon| < 1$ . In other words, no matter how many terms are there, the sum  $n * \epsilon = \epsilon + \epsilon + \epsilon + \cdots + \epsilon$  ( $n$  terms)  $< 1$ .

And in an Archimedean ordered field, the only infinitesimal is zero. When people become acquainted with the concept of limit, they know that “infinitesimal” should not be interpreted statically. Rather, it is the “vanishing behavior” of, e.g. the function  $(\sin^3 x)/x^2$ , when  $x \rightarrow 0$ , that is to be identified with “infinitesimal”. Along this direction there appeared several attempts to find (=construct) the system  $^*\mathbb{R}$  promised by Leibniz. We now describe a good and very natural try, just years before the discovery of Robinson, by Schmieden–Laugwitz.

**Definition 15 (Guiding Idea)** A nullifying sequence is considered as zero, a vanishing sequence is considered as an infinitesimal, and a constant sequence is identified with the number itself.

There may be better terminology, but here a sequence  $\langle x_n \rangle$  is called nullifying, if after a finite number of terms, the rest of them are all zero. And a sequence  $\langle x_n \rangle$  is called vanishing, if it is converging to zero.

**Definition 16 (Confusing!)** Limit and limiting-point recall the most confusing duo-definitions for a sophomore: “ $b$  is the limit of a sequence  $\langle a_n \rangle$ ”, and “ $c$  is a limiting point of a sequence  $\langle a_n \rangle$ ”. For any  $\epsilon > 0$ , there exists an  $m \in \mathbb{N}$ , such that,  $\forall n \geq m$ ,  $|a_n - b| < \epsilon$ . For any  $\epsilon > 0$ , and for any  $m \in \mathbb{N}$ , there exists an  $n \in \mathbb{N}$ , such that  $n > m$ , and  $|a_n - c| < \epsilon$ . (We support Kelley’s advocacy, replacing “limiting” by “cluster”).

**Definition 17 (Eventual Set, Frequent Set)** A subset  $A$  of  $\mathbb{N}$  is called “eventual”, if it is “cofinite” (=complement of a finite subset). That is:  $\mathbb{N} \setminus A$  is finite. It is called “frequent”, if it is infinite. If a property  $P(i)$  holds true for  $i$  be-

longing to an eventual set (the mention of which is unnecessary and omitted intentionally), we say  $P(i)$  holds true “for eventual  $i$ ”. Similar meaning is applied to the phrase “for frequent  $i$ ”. With this kinds of abuse of the English language, we may rephrase the definitions above as: For a sequence  $a := \langle a_n : n \in \mathbb{N} \rangle$ ,  $b \in \mathbb{R}$  is a limit of  $a$  iff, for any  $\epsilon > 0$ ,  $a_i \in (b - \epsilon \dots b + \epsilon)$  for eventual  $i$ ;  $c$  is a limiting-point of  $a$  iff, for any  $\epsilon > 0$ ,  $a_i \in (c - \epsilon \dots c + \epsilon)$  for frequent  $i$ .

**Remark 2** See p. 65 in [7], for the definitions of “eventually”, “frequently”, for a general directed set of indices. Here the (naturally) directed set  $\mathbb{N}$  is very special, because it is the least well-ordered infinite set.

**Notation 1 (Sequence Spaces)** We write  $x \in \mathbb{R}^{\mathbb{N}}$  to mean:  $x = \langle x_n \rangle$  is a sequence of real numbers. (This is in accordance to our notating the set of mappings). Here are some subsets of  $\mathbb{R}^{\mathbb{N}}$ , arranged decreasingly:

$\mathbb{R}_{\beta}^{\mathbb{N}}$  is the set of all bounded sequences;

$\mathbb{R}_c^{\mathbb{N}}$  is the set of all convergent sequences;

$\mathbb{R}_0^{\mathbb{N}}$  is the set of all sequences converging to zero (=vanishing);

$\mathbb{R}_\kappa^{\mathbb{N}}$  is the set of all eventually-zero (=nullifying) sequences.

**Definition 18 (Coordinatewise Operations on Sequence Spaces)**

$$\langle x_n \rangle + \langle y_n \rangle := \langle x_n + y_n \rangle;$$

$$\langle x_n \rangle - \langle y_n \rangle := \langle x_n - y_n \rangle;$$

$$\langle x_n \rangle * \langle y_n \rangle := \langle x_n * y_n \rangle.$$

**Lemma 2** The system  $\mathbb{R}^{\mathbb{N}}$  is a commutative unital ring, a real vector space, and thence a commutative algebra (with identity). The following inclusions are subalgebra embeddings:

$$\mathbb{R}_c^{\mathbb{N}} \subset \mathbb{R}_{\beta}^{\mathbb{N}} \subset \mathbb{R}^{\mathbb{N}},$$

Now according to the guiding idea, a number  $\lambda \in \mathbb{R}$  is identified with the sequence with all terms  $= \lambda$ . In this sense  $\mathbb{R}_{\beta}^{\mathbb{N}}$ , (and a fortiori  $\mathbb{R}^{\mathbb{N}}$ ), is an extension of  $\mathbb{R}$ , considered as an algebra. But this is not field-extension, because

**Example 1** Let  $1_E \in \mathbb{R}_{\beta}^{\mathbb{N}}$  and  $1_O \in \mathbb{R}_{\beta}^{\mathbb{N}}$  be the indicator sequences of the sets of even and odd indices, respectively. Then they are complementary idempotents:

$$\begin{aligned} 1_E * 1_E &= 1_E; & 1_O * 1_O &= 1_O; & 1_E * 1_O &= 0; \\ & & 1_E + 1_O &= 1. \end{aligned}$$

Since the two elements are neither 1, nor 0, of the algebra  $\mathbb{R}_{\beta}^{\mathbb{N}}$ , the latter cannot be a field, because even an 8th grader knows that in a field  $F$ , the only idempotent element  $x = x^2$  is  $x = 0$  (the zero) or  $x = 1$  (the identity).

**Definition 19 (Extension of Function)** Consider these two functions  $\cos$  and  $\sin$ . For a sequence  $x = \langle x_n \rangle \in \mathbb{R}^{\mathbb{N}}$ ,  $x$  is a “number” in a generalized sense, and we would like to define  $\cos(x)$ ,  $\sin(x)$ . If  $x$  is a constant sequence:  $x_n = a$ ,  $\forall n$ , then  $\cos(x)$ ,  $\sin(x)$ , should be the sequences with constant terms of  $\cos(a)$ ,  $\sin(a)$  correspondingly. So we see that a coordinatewise definition is almost forced on us: If  $f: \mathbb{R} \mapsto \mathbb{R}$  is everywhere defined, then we define:

$$\bar{f}\langle x_n \rangle := \langle f(x_n) \rangle.$$

So far as it is defined, it works fine. As an example: the addition-formula

$$\overline{\sin}(\langle x_n \rangle + \langle y_n \rangle) = \overline{\sin}\langle x_n \rangle * \overline{\cos}\langle y_n \rangle + \overline{\sin}\langle y_n \rangle * \overline{\cos}\langle x_n \rangle$$

would simply mean that, for each index  $n$ ,

$$\sin(x_n + y_n) = \sin(x_n) * \cos(y_n) + \sin(y_n) * \cos(x_n).$$

**Theorem 1 (A Transfer Principle: Restricted-Form)** If functions  $f, g, \dots$ , are all defined on  $\mathbb{R}$ , and they has a formula connecting them, then this formula still holds for the extended functions on the whole sequence space  $\mathbb{R}^{\mathbb{N}}$ .

We remark that coordinatewise extension is equally applicable to two-variable function, so long as it is defined everywhere. The Transfer Principle remains valid. And indeed the principle has nothing to do with the differentiability, continuity, etc., of the functions involved. This is pure naive-set theory!

**Remark 3** If bounded sequences are called “moderate numbers”, then only bounded functions will map “moderate number” to “moderate number”.

**Definition 20 (Ordering)** For two elements  $x = \langle x_n \rangle$ ,  $y = \langle y_n \rangle$  of  $\mathbb{R}^{\mathbb{N}}$ , we define:

$$(x \leq y), \quad \text{iff} \quad (\forall n, x_n \leq y_n).$$

We write  $x < y$ , iff  $x \leq y$ , and  $x \neq y$ . (This means the following:  $x_n \leq y_n$ , for all  $n \in \mathbb{N}$ , but there is an index  $n$ , (at least one, and only one is enough!) such that  $x_n < y_n$ ).

The binary relation  $\leq$  in  $\mathbb{R}^{\mathbb{N}}$  is called a partial ordering, because it is transitive, and antisymmetric: If  $x \leq y$ , and also  $y \leq x$ , then  $x = y$ . This partial ordering is not a total-ordering, because some pair of members, for example  $1_E$

and  $1_O$ , may be ‘incomparable’, as neither  $1_E \geq 1_O$  nor  $1_O \geq 1_E$  is true.

**Lemma 3 (Nullifying and Vanishing Sequences)** *The set  $\mathbb{R}_0^{\mathbb{N}}$  is an ideal of  $\mathbb{R}^{\mathbb{N}} \supset \mathbb{R}_b^{\mathbb{N}} \supset \mathbb{R}_c^{\mathbb{N}}$ . The set  $\mathbb{R}_c^{\mathbb{N}}$  is an ideal of  $\mathbb{R}_b^{\mathbb{N}} \supset \mathbb{R}_c^{\mathbb{N}}$ . Doing modulo  $\mathbb{R}_0^{\mathbb{N}}$  means that two sequences with vanishing difference are considered the same. Or, modulo  $\mathbb{R}_0^{\mathbb{N}}$  is a magnifying-glass not powerful enough to distinguish a vanishing sequence from zero sequence. And you don’t have non-trivial infinitesimals in the system. Also,*

**Lemma 4** *The quotient algebra  $\mathbb{R}_c^{\mathbb{N}} \bmod \mathbb{R}_0^{\mathbb{N}}$  is (isomorphic to)  $\mathbb{R}$ .*

The isomorphism identifies any convergent sequence  $\langle x_n \rangle$  with its limit  $\lim_{n \rightarrow \infty} x_n$ . The Lemma says that  $\lim_{n \rightarrow \infty} x_n * y_n = \lim_{n \rightarrow \infty} x_n * \lim_{n \rightarrow \infty} y_n$ , and the like. (Just the first theorems of our calculus text-book.)

**Definition 21 (Quotient by Eventuality)** We turn to the quotient algebra  $\mathbb{R}_{\mathbb{E}}^{\mathbb{N}} = \mathbb{R}^{\mathbb{N}} \bmod \mathbb{R}_c^{\mathbb{N}}$ , where two sequences  $x \in \mathbb{R}^{\mathbb{N}}$  and  $y \in \mathbb{R}^{\mathbb{N}}$  are equivalent mod  $\mathbb{R}_c^{\mathbb{N}}$ , iff  $x_n = y_n$  for eventual  $n$ . As we denote by  $\mathbb{E}$  the set of all eventual set of indices, we will also write  $x \sim_{\mathbb{E}} y$  or  $x \sim_{\mathbb{E}} y$  for this equivalence, and the equivalence class of  $x$  is written  $\pi_{\mathbb{E}}(x) \in \mathbb{R}_{\mathbb{E}}^{\mathbb{N}}$ . The system  $\mathbb{R}_{\mathbb{E}}^{\mathbb{N}}$  is a gentle modification of  $\mathbb{R}^{\mathbb{N}}$ . It has essentially the same good or bad properties of  $\mathbb{R}^{\mathbb{N}}$ . Because, in essence, the idea of  $\bmod \mathbb{E}$  is simply: When studying a sequence, never worry about its few exceptional terms! So we also can extend all the everywhere-defined functions for these “ $\mathbb{E}$ -numbers”. And the transfer principle still applies! Something better comes out from this modification  $\bmod \mathbb{E}$ : ordering.

**Definition 22 (Ordering)** For two elements  $x = \langle x_n \rangle$ ,  $y = \langle y_n \rangle$  of  $\mathbb{R}^{\mathbb{N}}$ , we write  $x \leq_{\mathbb{E}} y$ , and also  $\pi_{\mathbb{E}}(x) \leq \pi_{\mathbb{E}}(y)$ , iff

$$x_n \leq y_n \quad \text{for eventual } n.$$

Naturally we write  $x <_{\mathbb{E}} y$ , and also  $\pi_{\mathbb{E}}(x) < \pi_{\mathbb{E}}(y)$ , iff  $x \leq_{\mathbb{E}} y$ , and  $x \neq_{\mathbb{E}} y$ . This means:  $x_n \leq y_n$ , for eventual  $n$ , but  $x_n < y_n$ , for frequent  $n$ .

Note in the last sentence a subtle difference with the situation in  $\mathbb{R}^{\mathbb{N}}$ , where “for just one  $n$ ” will do. Again this partial ordering is not a total-ordering, as for example  $\pi_{\mathbb{E}}(1_E)$  and  $\pi_{\mathbb{E}}(1_O)$  are still “incomparable”. But the good news is there are “infinitesimals”!

**Definition 23** Let  $x \in \mathbb{R}^{\mathbb{N}}$ . If for all  $m \in \mathbb{N}$ ,  $\pi_{\mathbb{E}}(x) \leq \frac{1}{m}$ , then  $\pi_{\mathbb{E}}(x)$  is called an infinitesimal in  $\mathbb{R}_{\mathbb{E}}^{\mathbb{N}}$ .

**Lemma 5 (Infinitesimals)** *For  $x \in \mathbb{R}^{\mathbb{N}}$ ,  $\pi_{\mathbb{E}}(x)$  is an infinitesimal in  $\mathbb{R}_{\mathbb{E}}^{\mathbb{N}}$ , iff  $x \in \mathbb{R}_0^{\mathbb{N}}$ .*

If infinitesimals are just (the equivalence classes of) vanishing sequences, of course sequences “diverging to infinity” are then the infinities!

## Hyperreals

### The Construction of Ultimate-Reals

Let us review the system  $\mathbb{R}_{\mathbb{E}}^{\mathbb{N}}$ , especially the compatibility of the equivalence relation  $\sim_{\mathbb{E}}$  with multiplication:

$$((x \sim_{\mathbb{E}} y) \text{ and } (u \sim_{\mathbb{E}} v)) \Rightarrow (x * u) \sim_{\mathbb{E}} (y * v).$$

The proof is simple: the conditions  $x \sim_{\mathbb{E}} y, u \sim_{\mathbb{E}} v$  mean that  $A := \{i: x_i = y_i\} \in \mathbb{E}; B := \{i: u_i = v_i\} \in \mathbb{E}$ ; while  $A \cap B \subset \{i: x_i * u_i = y_i * v_i\}$ . And the compatibility condition is guaranteed by the obvious property that:

$$(ii): ((A \in \mathbb{E}) \& (B \in \mathbb{E})) \Rightarrow ((A \cap B) \in \mathbb{E}).$$

On the other hand, the appearance of the zero-divisors, (or equivalently, the nontrivial idempotents), corresponds to the fact that:  $(iv^{\times})$ : there is a set  $A$  of indices, such that both  $A$  and its complement  $B$  is not in  $\mathbb{E}$ . E. g.,  $A$ , and  $B$  are even and odd indices of  $\mathbb{N}$ . With these two remarks in background, we will now make a successful construction of the hyperreals. For the convenience of later possible generalizations, we will write  $I$  for  $\mathbb{N}$  as the set of indices. Indeed both the well-ordering and the countably infinite cardinality of  $\mathbb{N}$  are irrelevant here, and  $I$  is required to be an infinite set, just to make the construction interesting. But, for some deeper NSA, an index set of larger cardinality is really needed. So we first form the space  $\mathbb{R}^I$ , whose elements are real-valued functions on  $I$ . Such a function is still called a “sequence”, (even when  $I$  is uncountable), and we continue to write  $x(j)$  as  $x_j$ . As before, binary and unary operations on  $\mathbb{R}^I$  are defined coordinatewise. And in particular, it is a commutative algebra with identity.

**Definition 24 (Hyperreals)** Let  $\mathcal{U}$  be a free ultra-filter on  $I$ , i. e., a set of subsets of  $I$ , satisfying some conditions to be specified presently. We call two sequences  $x$  and  $y \in \mathbb{R}^I$   $\mathcal{U}$ -related, and we write  $x \sim_{\mathcal{U}} y$ , iff:

$$\{j \in I: x_j = y_j\} \in \mathcal{U}.$$

The equivalence class of  $x$  is denoted by  $\pi_{\mathcal{U}}(x)$  or  $x \bmod (\sim_{\mathcal{U}})$ , and called a hyperreal (along  $\mathcal{U}$ ). The quotient set will be denoted simply as  ${}^*\mathbb{R}$ ,  $\mathcal{U}$  being understood. The canonical mapping from  $\mathbb{R}^I$  will be denoted by  $\pi_{\mathcal{U}}$ .

**Definition 25 (Ultra-Filter)** A subset  $\mathcal{U}$  of  $\mathfrak{P}(I)$  is called a free-ultrafilter on the infinite set  $I$ , if the following conditions are fulfilled:



- (i'): if  $A \in \mathcal{U}$  then  $A$  is non-empty,
- (i): if  $A \in \mathcal{U}$  then  $A$  is not a finite set;
- (ii): if  $A \in \mathcal{U}$  and  $B \in \mathcal{U}$ , then  $A \cap B \in \mathcal{U}$ ;
- (iii): if  $A \in \mathcal{U}$  and  $B \supset A$ , then  $B \in \mathcal{U}$ ;
- (iv): if  $A \subset I$ , then either  $A \in \mathcal{U}$  or  $(I \setminus A) \in \mathcal{U}$ .

Remark: We may say that the information about an element  $x \in \mathbb{R}^I$  is indexed by  $I$ . If  $A \in \mathcal{U}$ , then the information about this equivalence class  $\pi_{\mathcal{U}}(x)x$  is completely contained in  $\{x_j: j \in A\}$ , the information about  $x$  on that  $A$ -indexed part, while the part off  $A$  is irrelevant for all(our) purpose. (iii) means: "more than enough (part) is enough". (ii) means: "if both  $A$  part and  $B$  part are enough, actually the intersection part is enough". In mathematical language, if  $I$  is any (infinite) set, and  $\mathcal{U} \subset \mathfrak{P}(I)$ ,  $\mathcal{U}$  is called a filter on  $I$ , if it satisfies the condition (i', ii, iii). If the condition (i) also holds, the filter  $\mathcal{U}$  is called free. The dichotomy condition (iv) for an ultra-filter is very strange: "if the information about  $A$  part and the information about  $B$  part, when gathered jointly, is enough, then actually one or the other part must be enough".

*Note 2 (Non-free Ultrafilter and Axiom of Choice)* A non-free ultra-filter is trivial because it is then the set of all sets of indices containing a particular index  $i_0$ . In that case,  $x \stackrel{\mathcal{U}}{\sim} y$  would simply mean  $x(i_0) = y(i_0)$ . In other words: the only relevant information is this  $i_0$  item. The whole indexing of information would be non-sense: there is no need of votes-counting, because the policy is determined by the great-leader uniquely. The existence of a free ultra-filter is guaranteed by the axiom of choice. (Indeed we are told that this is a weaker form of the axiom choice).

From now on, a free ultra-filter  $\mathcal{U}$  is chosen. Again by abuse of English, a set  $A \in \mathcal{U}$  is called an ultimate (set of indices). And "for ultimate  $j$ " means that set of indices  $j$  validating the assertion concerned is in  $\mathcal{U}$ . Also, the complement of an ultimate, or by the dichotomy equivalently a non-ultimate, is called an evanescent (or "negligible") set of indices.

**Theorem 2 (Totally Ordered Field)** *The quotient system  ${}^*\mathbb{R}$  of the algebra  $\mathbb{R}^I$  by the equivalence relation  $\stackrel{\mathcal{U}}{\sim}$ , is a totally ordered field.*

Proof of the compatibility of the operations and  $\stackrel{\mathcal{U}}{\sim}$ : These are all tedious, but they are easy consequences of the filtering condition ii. We will take division as an example. Let  $\pi_{\mathcal{U}}(x) = \pi_{\mathcal{U}}(u) = \xi \in {}^*\mathbb{R}$ ,  $\pi_{\mathcal{U}}(y) = \pi_{\mathcal{U}}(v) = \eta \in {}^*\mathbb{R}$ , so that  $x \stackrel{\mathcal{U}}{\sim} u$ ,  $y \stackrel{\mathcal{U}}{\sim} v$ , and we have to show that

$$\left\{ j \in I: \frac{x_j}{y_j} = \frac{u_j}{v_j} \right\} \in \mathcal{U},$$

and then  $\frac{\xi}{\eta} \in {}^*\mathbb{R}$  is unambiguously defined by

$$\frac{\xi}{\eta} := \pi_{\mathcal{U}} \left\langle \frac{x_j}{y_j} \right\rangle.$$

Of course it is assumed that (in  ${}^*\mathbb{R}$ ),  $\eta \neq 0$ . That is

$$A := \{j \in I: y_j \neq 0\} \in \mathcal{U}.$$

Now  $\pi_{\mathcal{U}}(x) = \pi_{\mathcal{U}}(u)$ ,  $\pi_{\mathcal{U}}(y) = \pi_{\mathcal{U}}(v)$ , we have:

$$B := \{j \in I: x_j = u_j\} \in \mathcal{U},$$

$$C := \{j \in I: y_j = v_j\} \in \mathcal{U}.$$

And thus  $D := A \cap B \cap C \in \mathcal{U}$ . For index  $j \in D$ , we have:

$$\frac{x_j}{y_j} = \frac{u_j}{v_j},$$

which is what to be shown. Caution: For  $j \notin A$ ,  $x_j/y_j$  is not defined! The obvious remedy is thus: a hyperreal  $\eta \in {}^*\mathbb{R}$  is represented by a "sequence"  $y = (y_j: j \in A)$ , which is defined by an ultimate set  $A \in \mathcal{U}$ .

Proof of the total ordering: For  $x \in \mathbb{R}^I$ ,  $y \in \mathbb{R}^I$ , consider the sets  $A := \{j: x_j \geq y_j\}$ ,  $B := \{j: y_j \geq x_j\}$ . Now  $A \cup B = I \in \mathcal{U}$ , so either  $A \in \mathcal{U}$ , or  $B \in \mathcal{U}$ , by dichotomy. Thus  $\pi_{\mathcal{U}}(x) \geq \pi_{\mathcal{U}}(y)$  or  $\pi_{\mathcal{U}}(y) \geq \pi_{\mathcal{U}}(x)$ . Needless to say: when  $A \in \mathcal{U}$ , and  $B \in \mathcal{U}$ , then  $A \cap B = \{j: x_j = y_j\} \in \mathcal{U}$ , thus  $x \stackrel{\mathcal{U}}{\sim} y$ . Note that in the proof above for division, we actually proved the multiplicative invertibility for  $\eta \neq 0$ , therefore we have shown that in

${}^*\mathbb{R}$

there is no non-trivial idempotent. as there is no zero-divisors.

**Definition 26 (Infinitesimals and Infinities)** Since  $\mathbb{R}$  is embedded canonically in the totally ordered field  ${}^*\mathbb{R}$ , a hyperreal  $\alpha$  is called infinitesimal iff: for any  $n \in \mathbb{N}$ ,  $|\alpha| < \frac{1}{n}$ . The set of all infinitesimals is denoted by  $\mathbb{I}_0$ , while the subset of all non-zero infinitesimals is denoted by  $\mathbb{I}$ .

The set  $\mathbb{I}_0$  is a real linear subspace of  ${}^*\mathbb{R}$ . Note that the only standard member therein is zero. So  $\mathbb{I} \cap \mathbb{R} = \emptyset$ .

**Definition 27 (Monad)** Any two hyperreals with infinitesimal difference are called infinitely close. This is an equivalence relation of "infinitely close", written  $\approx$ , whose equivalence classes are called "monads", after Leibniz.

**Definition 28 (Limited and Unlimited)** A hyperreal  $\alpha \in {}^*\mathbb{R}$  is called limited (= "finite") if, for some  $n \in \mathbb{N}$ ,  $-n \leq \alpha \leq n$ . In this case, there is uniquely one real number  $a = {}^\circ\alpha$  infinitely close to  $\alpha$ . Then  $\alpha = a + (\alpha - a)$ ,  $a$  is

the standard part of  $\alpha$ , and  $\alpha - a$  is the infinitesimal part of  $\alpha$ . On the other hand, a hyperreal  $\alpha$  is called limited (“infinite”) iff: for any  $n \in \mathbb{N}$ ,  $|\alpha| > n$ .

An equivalent definition of this infinity is that:  $\frac{1}{\alpha} \in \mathbb{I}$ . The set of all limited hyperreals is a subalgebra (with identity) of  ${}^*\mathbb{R}$ , with the set  $\mathbb{I}_0$  of infinitesimals as an ideal. The quotient algebra is isomorphic with the real field  $\mathbb{R}$ , so the set of all limited hyperreals is indeed

$$\mathbb{R} + \mathbb{I}_0 = \{c + \zeta : c \in \mathbb{R}, \zeta \in \mathbb{I}_0\}.$$

Note: Keeping  $I = \mathbb{N}$ , and  $O, E$  the subsets of odd and even indices respectively, one of the two sequences  $1_O, 1_E$ , is (equivalent to) the unit ( $= 1$ ) of  ${}^*\mathbb{R}$ , while the other is the zero ( $= 0$ ) of  ${}^*\mathbb{R}$ . (They are idempotents complementing each other). The choice is made by the ultrafilter  $\mathcal{U}$ . Or, we can choose one or the other from the two sets  $O, E$ , to be a member of  $\mathcal{U}$ . The axiom of choice allow us to do this before the final set  $\mathcal{U}$  is ultimately found.

*Example 2 (A monad is at least a continuum)* When  $I = \mathbb{N}$ , the hyperreal  $\varepsilon = \pi_{\mathcal{U}}(\frac{1}{n} : n \in \mathbb{N})$  is infinitesimal. Then so is  $\varepsilon^3$ , and indeed for any positive real number  $r > 0$ ,  $\varepsilon^r$  is infinitesimal. These  $\varepsilon^r$  are inversely ordered as  $r$ . This is easily seen from the representing sequences  $\langle n^{-r} : n \in \mathbb{N} \rangle$ . So the monad of infinitesimals is at least a continuum, cardinally speaking.

*Example 3 (A smaller infinitesimal)* Suppose real numbers  $r > 1, s > 1$  are fixed. Then for eventual  $n \in \mathbb{N} = I$ ,

$$s^{-n} < \frac{1}{n^r};$$

A fortiori, this is true for ultimate  $n$ ; therefore

$$\nu_s = \pi_{\mathcal{U}}(s^{-n} : n \in \mathbb{N}) < \varepsilon^r.$$

To an atheist-Buddhist, an ultrafilter  $\mathcal{U}$  is a Buddha’s magnifying glasses, viewing through which a tiny monad is bigger than a solar system.

### Star-Extensions

Now we come to the central issue of function-extension and the intimately related transfer principle. For an everywhere-defined function  $f$ , the coordinatewise definition still works fine, and we surely have an extension  $\bar{f}$  of  $f$  to the hyperreals. (The same proof works. Or rather, the assertion here is then a corollary of the former principle.) For a function like  $\text{csc}$ , which is undefined at  $n\pi \in \pi * \mathbb{Z}$ ,  $\overline{\text{csc}}$  as definable at  $x = \langle x_n \rangle$ , is

$$\overline{\text{csc}}(x) := \langle \text{csc}(x_n) \rangle \in \mathbb{R}^{\mathbb{N}}.$$

In  $\mathbb{R}^{\mathbb{N}}$ , the definition breaks down, if only one term  $x_n \in \pi * \mathbb{Z}$ . How about the situation in  $\mathbb{R}_{\mathcal{G}}^{\mathbb{N}}$ ? Write temporarily the domain of definition of  $\text{csc}$  as  $A := \text{Dom}(\text{csc}) = \mathbb{R} \setminus \pi * \mathbb{Z}$ , then if for eventual  $n$ ,  $x_n \in B$ , the definition goes through! And for such an  $x \in \mathbb{R}^{\mathbb{N}}$ ,  $\overline{\text{csc}}(\pi_{\mathcal{G}}(x)) \in \mathbb{R}_{\mathcal{G}}^{\mathbb{N}}$  is defined. In complete analogy, for an  $x \in \mathbb{R}^{\mathbb{N}}$ , if only: “for ultimate  $n$ ,  $x_n \in A$ ”,  $\pi_{\mathcal{U}}(\langle \text{csc}(x_n) \rangle) \in {}^*\mathbb{R}$  is well-defined!

**Definition 29 (Star-Extension of a Set)** For a non-empty set  $A \subset \mathbb{R}$ , the set  ${}^*A$  is the set of all  $\pi_{\mathcal{U}}(x)$ , where  $x = \langle x_j \rangle$  is a “sequence” such that, for ultimate  $j$ ,  $x_j \in A$ .

**Definition 30 (Extension of a Mapping)** For any mapping  $f: A \mapsto B$ , and  $x = \langle x_j \rangle \in A^I$ , we have the “sequence”  $y = f \circ x = \langle f(x_j) \rangle \in B^I$ , whose equivalence class  $\eta = \pi_{\mathcal{U}}(y) \in {}^*B$  is unambiguously determined by  $\xi = \pi_{\mathcal{U}}(x) \in {}^*A$ , and is independent of the representative  $x \in A^I$ . We then write  $\eta = {}^*f(\xi)$ . This mapping  ${}^*f: {}^*A \mapsto {}^*B$  is called the star-extension of  $f$ . (It is surely an extension.)

It is of utmost importance to note here that these definitions of “star-extensions” are strictly notions of naive set theory. Of course  $A, B$  can be any two non-empty sets, and  $f$  any function from  $A$  to  $B$ .

**Theorem 3 (Transfer Principle)** All the identities for the standard functions hold true for their star-extensions, simply because they hold true for the original functions.

*Example 4 (HyperIntegers)* A hyperinteger  $\pi_{\mathcal{U}}(x) \in {}^*\mathbb{Z}$  is the equivalence class of a sequence of integers  $x \in \mathbb{Z}^{\mathbb{N}}$ . If “integer” is changed to “even integer”, then we get even hyperinteger  $\pi_{\mathcal{U}}(x) \in {}^*(2 * \mathbb{Z})$ , where  $2 * \mathbb{Z} := \{2 * n : n \in \mathbb{Z}\}$  is the set of even integers. And there are odd hyperintegers, too. All the hyperintegers may be classified as even or odd, because of the dichotomy property of the ultrafilter  $\mathcal{U}$ .

*Example 5 (Cartesian Product and Two Variable Function)* For two non-empty sets  $A, B$ , by canonical identification,

$${}^*(A \times B) = {}^*A \times {}^*B.$$

So for any mapping  $g: A \times B \mapsto C$ , the star-extension of  $g$  may be defined:  ${}^*g: {}^*A \times {}^*B \mapsto {}^*C$ .

What this says is easy but tedious to explain. First, we know the identification:

$$(A \times B)^I = A^I \times B^I.$$

Because for  $x \in A^I, u \in B^I$ , the couple  $\langle x, u \rangle \in A^I \times B^I$  corresponds bijectively to an element  $x \otimes u$  of  $(A \times B)^I$ ,

by:

$$x \otimes u : i \in I \mapsto \langle x_j, u_j \rangle \in A \times B;$$

And the equivalence class of  $x \otimes u$  depends on the equivalence classes  $\pi_{\mathcal{U}}(x)$  and  $\pi_{\mathcal{U}}(u)$  only. In this way we are sure that  $\pi_{\mathcal{U}}(g(x_j, y_j))$  is unambiguously determined by  $\pi_{\mathcal{U}}(x)$  and  $\pi_{\mathcal{U}}(u)$  only, and is independent of the representative  $x \in A^I, y \in B^I$ . Actually we did give a proof for  $g = \text{division}$ ,  $A = \mathbb{R} = C, B = \mathbb{R}^\times = \mathbb{R} \setminus \{0\}$ . The proof is good for any function of any (finite) number of variables. Some examples of Transfer are:

- $^*\text{abs}(\alpha) = |\alpha| \in {}^*\mathbb{R}$  is meaningful for  $\alpha \in {}^*\mathbb{R}$ . And indeed  $|\alpha| = -\alpha$ , if  $\alpha < 0$ .
- $(-1)^\mu$  is meaningful for  $\mu \in {}^*\mathbb{Z}$ . And indeed it is 1 for  $\mu$  an even hyperinteger, it is  $-1$  for  $\mu$  an odd hyperinteger.
- The star-extensions  $^*\sin, ^*\cos$  of the sine, cosine functions are defined on  ${}^*\mathbb{R}$  and valued in the star-extended unit closed interval  ${}^*[0 \dots 1]$ .

Note: From now on, the star-extensions of such classical functions like  $\sin, \cos, \dots$ , will not have the star displayed.

### NS View of Continuous Functions

The introduction of hyperreals gives us a different standpoint to view many familiar concepts.

*Remark 4 (Extended Version of a Sequence)* Suppose  $s = \langle s_n : n \in \mathbb{N} \rangle$  is a sequence of real numbers. This is a function from  $\mathbb{N}$  to  $\mathbb{R}$ . Therefore, it may be star-extended to a function

$$^*s : m \in {}^*\mathbb{N} \mapsto s_m \in {}^*\mathbb{R}.$$

We call this latter the extended version of the original sequence. Indeed for any unlimited integer  $v = \pi_{\mathcal{U}}(n_j : j \in \mathbb{N})$ , we may define

$$s_v := \pi_{\mathcal{U}}(s_{n_j} : j \in \mathbb{N}) \in {}^*\mathbb{R}.$$

It is obvious that the convergence of the sequence  $s$

$$\lim_{n \rightarrow \infty} s_n = b,$$

is equivalent to:

$$\text{for all unlimited integer } v \in {}^*\mathbb{N} \setminus \mathbb{N}, \quad s_v \approx b.$$

As an example, when  $r \in \mathbb{R}, y \in \mathbb{R}$ , and

$$s_n := \cos^m \left( \frac{y}{\sqrt{n}} \right), \quad \text{with } m = \text{floor}(n * r),$$

standard calculus gives us

$$\lim_{n \rightarrow \infty} s_n = e^{-\frac{y^2 r}{2}}.$$

We will talk about a function  $f : \mathbb{R} \mapsto \mathbb{R}$ , though usually the domain of definition is only required to be an open interval.

**Theorem 4** A function  $f$  is bounded around a point  $a \in \mathbb{R}$ , iff  ${}^*f$  maps the monad of  $a$  into finite hyperreals.

*Proof* If  $f$  maps a neighborhood  $(a - e \dots a + e)$  of  $a$  into  $[-M \dots M]$ , where  $e \in \mathbb{R}$  is positive, then for  $\xi = \pi_{\mathcal{U}}(x_j) \approx a$ , we have  $a - e < x_j < a + e$  for ultimate  $j$ , and thence  $f(x_j) \in [-M \dots M]$ . Accordingly:

$$-M^* \leq {}^*f(\xi)^* \leq M.$$

On the other hand, if  $f$  is unbounded around  $a$ , then for any  $j \in I = \mathbb{N}$ , there is  $x_j$  (in the domain of definition of  $f$ ), with  $|x_j - a| < \frac{1}{j}$ , and  $f(x_j) > j$ . Thus  $\xi := \pi_{\mathcal{U}}(x_j) \approx a$ , but  $|{}^*f(\xi)|$  is infinite.  $\square$

**Theorem 5** Let  $A \subset \mathbb{R}$  be an interval,  $b \in A$ , and  $f : A \mapsto \mathbb{R}$ . Then  $f$  is continuous at  $b$ , iff: for all  $\xi \in {}^*A$ ,

$$\xi \approx b \Rightarrow {}^*f(\xi) \approx f(b).$$

In other words,  $f$  is continuous at  $b$ , iff  ${}^*f$  maps the monad  $a + \mathbb{I}_0$  into the monad  $f(a) + \mathbb{I}_0$ .

*Proof* By the standard definition, when  $f$  is continuous at  $b$ , and  $\epsilon > 0$  is given, there is a  $\delta > 0$ , such that, for  $|x - b| < \delta$ , (and  $x \in A$ ),  $|f(x) - f(b)| < \epsilon$ . Now  $\xi = \pi_{\mathcal{U}}(x_j)$ , and  $\xi - b \in \mathbb{I}_0$ , so for ultimate  $j$ ,  $|x_j - b| < \delta$ . Thus  $|f(x_j) - f(b)| < \epsilon$ , for ultimate  $j$ . This means  ${}^*f(\xi) - f(b) \in \mathbb{I}_0$ , or,  ${}^*f(\xi) \approx f(b)$ . Suppose  $f$  is not continuous at  $b$ , then there is  $\epsilon > 0$ , such that for whatever  $j \in \mathbb{N}$ , there is  $x_j \in A$ , with  $|x_j - b| < \frac{1}{j}$ , while  $|f(x_j) - f(b)| > \epsilon$ . Then with  $\xi = \pi_{\mathcal{U}}(x_j) \in {}^*A, \xi \approx b$ , while  $|{}^*f(\xi) - f(b)| > \epsilon$ .  $\square$

**Corollary 1** If  $f$  is continuous at  $c$ , it is bounded around  $c$ .

**Theorem 6** Let  $A \subset \mathbb{R}$  be a compact (=closed and bounded) interval, and  $f : A \mapsto \mathbb{R}$  is continuous. Then  $f$  is bounded.

*Proof* If  $f$  is unbounded, then for  $j \in \mathbb{N}$ , there is  $x_j \in A$ , with  $|f(x_j)| > j$ . And we find  $\xi = \pi_{\mathcal{U}}(x_j)$ , with  ${}^*f(\xi)$  infinite. But  $\xi \in {}^*A$  is a limited hyperreal, and its standard part  $b = {}^\circ\xi \in A, b \approx \xi$ . By the continuity criterion,  ${}^*f(\xi) \approx f(b) \in \mathbb{R}$ . This is a contradiction.  $\square$

**Theorem 7** Let  $A \subset \mathbb{R}$  be an interval, and  $f : A \mapsto \mathbb{R}$ . Then  $f$  is uniformly continuous, iff: for all  $\xi \in {}^*A, \eta \in {}^*A$ ,

$$\xi \approx \eta \Rightarrow {}^*f(\xi) \approx {}^*f(\eta).$$

*Proof* By the standard definition, when  $f$  is uniformly continuous on  $A$ , and  $\epsilon > 0$  is given, there is a  $\delta > 0$ , such that, for  $|x - y| < \delta$ , (and  $x \in A, y \in A$ ),  $|f(x) - f(y)| < \epsilon$ . Now  $\xi = \pi_{\mathbb{U}}\langle x_j \rangle \in {}^*A$ ,  $\eta = \pi_{\mathbb{U}}\langle y_j \rangle \in {}^*A$ , and  $\xi - \eta \in \mathbb{I}_0$ , so for ultimately many  $j$ ,  $|x_j - y_j| < \delta$ . Thus  $|f(x_j) - f(y_j)| < \epsilon$ , for ultimately many  $j$ . This means  ${}^*f(\xi) - {}^*f(\eta) \in \mathbb{I}_0$ , or,  ${}^*f(\xi) \approx {}^*f(\eta)$ . Suppose  $f$  is not uniformly continuous, then there is  $\epsilon > 0$ , such that for whatever  $j \in \mathbb{N}$ , there is a pair  $x_j \in A, y_j \in A$ , with  $|x_j - y_j| < \frac{1}{j}$ , while  $|f(x_j) - f(y_j)| > \epsilon$ . Then with  $\xi = \pi_{\mathbb{U}}\langle x_j : j \rangle \in {}^*A$ ,  $\eta = \pi_{\mathbb{U}}\langle y_j : j \in I \rangle \in {}^*A$ ,  $\xi \approx \eta$ , while  $|{}^*f(\xi) - {}^*f(\eta)| > \epsilon$ . Contradiction. One of the merits of NSA is that: a standard notion may be interpreted in a nonstandard way, which is pedagogically advantageous. There will be, as examples, a couple of theorems easily proved by the NS criterion just given. But first we introduce some notions very useful later on.  $\square$

**Definition 31 (Hyperfinite Set)** Let  $T_j \subset A$  for each  $j \in I$ . Then the sequence defines an “internal” subset  $\mathcal{T} := \pi_{\mathbb{U}}(T)$  of  ${}^*A$ :

$$\pi_{\mathbb{U}}(T) := \{\pi_{\mathbb{U}}(x) : x_j \in T_j \text{ for ultimate } j\} \subset {}^*A.$$

If for ultimate  $j$ , the set  $T_j$  is finite:  $\text{card}(T_j) < \infty$ ,  $\mathcal{T}$  is called a hyperfinite set with internal cardinal  $\pi_{\mathbb{U}}(\text{card}(T_j)) \in {}^*\mathbb{N}$ .

*Example 6 (Hyperfinite Equi-Partition)* Consider the hyperfinite set  $\mathcal{T}$  of equi-partition points with cardinal  $1 + \mathcal{N}$ :

$$\mathcal{N} := \pi_{\mathbb{U}}\langle N_j \rangle \in \mathbb{N}^I; (N_j < N_{j+1} \rightarrow \infty);$$

$$T_j := \left\{ \frac{k}{N_j} : k \in \mathbb{Z}, 0 \leq k \leq N_j \right\};$$

$$\mathcal{T} := \pi_{\mathbb{U}}\langle T_j \rangle.$$

Three convenient choices are  $N_j = j!$  and  $N_j = 2^j$ , and  $N_j = 3^j$ . The corresponding partition set  $\mathcal{T}$  is then called, respectively, the partition by factorial, or binary, or ternary fractions.

**Lemma 6 (Density of Equi-Partition Set)** *The countable set*

$$T_{\infty} = \cup_j T_j$$

*is dense in the unit interval  $\mathbb{U}$ : For any number  $c \in \mathbb{U}$ , there is a hyperreal  $\zeta \in {}^*T$ ,  $\zeta \approx c$ .*

*Proof* There are two convenient choices: the left- and the right-approximation:

$$\zeta = \text{floor}_{\mathcal{T}}(c) = \pi_{\mathbb{U}}\left\langle \frac{\text{floor}(c * N_j)}{N_j} : j \in I \right\rangle;$$

$$\text{or } \zeta = \text{ceil}_{\mathcal{T}}(c) = \pi_{\mathbb{U}}\left\langle \frac{\text{ceil}(c * N_j)}{N_j} : j \in I \right\rangle.$$

$\square$

Note that, for any  $x \in \mathbb{U}$ ,

$$\text{st}(\text{floor}_{\mathcal{T}}(x)) = x, \quad \text{st}(\text{ceil}_{\mathcal{T}}(x)) = x.$$

Both  $\text{floor}_{\mathcal{T}}$  and  $\text{ceil}_{\mathcal{T}}$  are (Skolem?) “inverse- function” of the standard-part mapping  $\text{st}$  from  $\mathbb{U}$  onto  $\mathcal{T}$ . Indeed for the binary-rational partition  $T_j = 2^j$ , or the ternary-rational partition  $T_j = 3^j$ , the floor-approximation gives the binary or ternary expansion respectively.

**Theorem 8** *Let  $A \subset \mathbb{R}$  be a compact interval, and  $f : A \rightarrow \mathbb{R}$  is continuous. Then  $f$  has a maximum and minimum.*

*Proof* Now (by simple scaling and translations) we may and will assume  $A = \mathbb{U} = [0 \dots 1]$  the unit interval. Take a hyperinteger  $\mathcal{N} := \pi_{\mathbb{U}}\langle N_j \rangle$ , and then the equi-partition  $\mathcal{T} = \pi_{\mathbb{U}}(T)$  into  $\mathcal{N}$  parts. Here  $T_j := \{k/N_j : k \in \mathbb{Z}, 0 \leq k \leq N_j\}$ . Now for each  $j$ , the maximum value of

$$f[T_j] := \{f(x) : x \in T_j\}$$

is unique, though the maximum point  $x_j$  may be non-unique:

$$f(x_j) \geq f(x), \quad \forall x \in T_j;$$

We just choose (say) the least one such maximum point  $x_j$ . And define

$$\xi := \pi_{\mathbb{U}}(x) \in {}^*A; \quad x_0 := {}^\circ \xi; \quad x_0 \in A.$$

Then for any  $c \in A$ , and  $\zeta$  as before,

$${}^*f(\zeta) \leq {}^*f(\xi); \quad \zeta \approx c;$$

$$x_0 \approx \xi; \quad {}^*f(x_0) \approx {}^*f(\xi); \quad {}^*f(c) \approx {}^*f(\zeta);$$

We see that  $f(c) \leq f(x_0)$ ,  $\forall c \in A$ , i. e.,  $f$  attains its maximum-value at  $x_0$ .  $\square$

**Theorem 9 (Intermediate-Value)** *Let  $f : A \rightarrow \mathbb{R}$  be continuous,  $v \in \mathbb{R}$ ,  $A = [a \dots b]$ , and  $(v - f(a))(v - f(b)) < 0$ . Then there is  $u \in (a \dots b)$ , such that  $f(u) = v$ .*

*Proof* We may and will assume  $A = \mathbb{U}$ ,  $a = 0$ ,  $b = 1$ ,  $f(0) < 0 < f(1)$ . As before, we take a hyper-finite  $\mathcal{N}$

equi-partition  $\mathcal{T} = \pi_{\mathbb{U}}\langle T_j \rangle$ ,  $\mathcal{N} = \langle N_j \rangle$ . Find the least  $x_j \in T_j$ , such that  $f(x_j) \geq 0$ . Because  $0 \in T_j$ ,  $1 \in T_j$ , and  $f(0) < 0 < f(1)$ , we know  $f(x_j) \geq 0$ ,  $f(x_j - 1/N_j) < 0$ . With  $\xi := \pi_{\mathbb{U}}\langle x_j \rangle \in {}^*A$ ,  $u := {}^\circ\xi$ , then  $\epsilon := \pi_{\mathbb{U}}\langle 1/N_j \rangle$ , we have:

$${}^*f(\xi) \geq 0, \quad {}^*f(\xi - \epsilon) < 0.$$

But  $\xi \approx u$ ;  $\xi - \epsilon \approx u$ ; and the continuity of  $f$  at  $u$  implies:

$${}^*f(\xi) - f(u) \in \mathbb{I}_0, \quad {}^*f(\xi - \epsilon) - f(u) \in \mathbb{I}_0.$$

Therefore:

$$\begin{aligned} 0 &\leq {}^*f(\xi) < {}^*f(\xi) - {}^*f(\xi - \epsilon) \\ &= ({}^*f(\xi) - f(u)) - ({}^*f(\xi - \epsilon) - f(u)) \in \mathbb{I}_0, \end{aligned}$$

resulting in  ${}^*f(\xi) \in \mathbb{I}_0$ . Or  ${}^\circ({}^*f(\xi)) = 0$ , which means  $f(u) = 0$ .  $\square$

**Theorem 10** Let  $A \subset \mathbb{R}$  be a compact interval, and  $f: A \mapsto \mathbb{R}$  be continuous. Then  $f$  is uniformly continuous.

*Proof* Suppose  $\xi \in {}^*A$ ,  $\eta \in {}^*A$ , and  $\xi \approx \eta$ . Now let  $c = {}^\circ(\xi) \in A$ . It follows that both  $\xi \approx c$  and  $\eta \approx c$  hold true. The continuity of  $f$  implies that both  ${}^*f(\xi) \approx f(c)$  and  ${}^*f(\eta) \approx f(c)$  hold. Therefore  ${}^*f(\xi) \approx {}^*f(\eta)$ .  $\square$

## NS Calculus

According to the standard definition:

$$f'(a) = \lim_{\Delta x \rightarrow 0} \frac{f(a + \Delta x) - f(a)}{\Delta x};$$

the differentiability of  $f$  at  $a$  implies the continuity of  $f$  at  $a$ , which means that  ${}^*f$  maps the monad  $a + \mathbb{I}_0$  of  $a$  into the monad  $f(a) + \mathbb{I}_0$  of  $f(a)$ .

**Definition 32 (Difference)** If  $f$  is a real-valued function defined around  $a$ , then the “discern” of  $f$  at  $a$  is the (non-standard) function

$$\Delta_a f: dx \in \mathbb{I}_0 \mapsto {}^*f(a + dx) - f(a) \in {}^*\mathbb{R}.$$

Of course we know that the continuity of  $f$  at  $a$  means that the range of  $\Delta_a f$  is included in  $\mathbb{I}_0$ . So we are led to consider (non-standard) functions from  $\mathbb{I}_0$  to  $\mathbb{I}_0$ . From now on,  $dx, dy, \dots$  will be variables whose values are infinitesimals. (So they live in  $\mathbb{I}_0$ .)

**Definition 33 (Tangential and Tangent)** If  $\Phi: \mathbb{I}_0 \mapsto \mathbb{I}_0$ ,  $\Psi: \mathbb{I}_0 \mapsto \mathbb{I}_0$ , and for any nonzero infinitesimal  $\epsilon \in \mathbb{I}$ ,

$$\frac{\Phi(\epsilon) - \Psi(\epsilon)}{\epsilon} \approx 0,$$

they are called tangential to each other. For any number  $c \in \mathbb{R}$ , the associated tangent is the (non-standard) function:

$$dx \in \mathbb{I}_0 \mapsto c * dx \in \mathbb{I}_0.$$

**Theorem 11 (Differentiability)** A function  $f$  defined on an open interval  $U$  is differentiable at  $a \in U$ , and there with derivative  $f'(a) = c$ , iff:  $\Delta_a f$  is tangential to the tangent associated with  $c$ .

*Proof* If  $f'(a) = c$ , then for any  $k \in \mathbb{N}$ , there is  $e_k \in \mathbb{R}$  with  $e_k > 0$ , such that whenever  $|u| < e_k$ ,  $|(f(a + u) - f(a))/u - c| < \frac{1}{k}$ . Now fix  $k \in \mathbb{N}$ . Thus to any (non-zero)  $dx = \pi_{\mathbb{U}}\langle u_j \rangle \in \mathbb{I}$ , for ultimate  $j$ ,  $0 < |u_j| < e_k$ , resulting in  $|(f(a + u_j) - f(a))/u_j - c| < \frac{1}{k}$ . So

$$\left| \frac{\Delta_a f(dx) - c * dx}{dx} \right| = \left| \frac{{}^*f(a + dx) - f(a)}{dx} - c \right| < \frac{1}{k}.$$

This being true for all  $k \in \mathbb{N}$ , we have: for any  $dx \in \mathbb{I}$ ,

$$\left| \frac{\Delta_a f(dx) - c * dx}{dx} \right| \in \mathbb{I}_0;$$

which means:  $\Delta_a f$  is tangential to the tangent associated with  $c$ . Conversely, if  $f$  is not differentiable at  $a$ , one can show that the nonstandard condition of differentiability can not be satisfied. Indeed then the following four limits of the difference quotient  $\frac{f(a+u)-f(a)}{u}$  are not all finite and coincident:

$$\limsup_{u \downarrow 0}, \quad \liminf_{u \downarrow 0}, \quad \limsup_{u \uparrow 0}, \quad \liminf_{u \uparrow 0}.$$

For example, a case may be

$$\begin{aligned} b &:= \limsup_{u \downarrow 0} \frac{f(a + u) - f(a)}{u} \\ &> e := \liminf_{u \downarrow 0} \frac{f(a + u) - f(a)}{u}. \end{aligned}$$

In that case, there are positive sequences  $u \in \mathbb{R}^{\mathbb{N}}$ ,  $v \in \mathbb{R}^{\mathbb{N}}$ , decreasing to 0, such that

$$\begin{aligned} b &:= \lim_{j \rightarrow \infty} \frac{f(a + u_j) - f(a)}{u_j}, \\ e &:= \lim_{j \rightarrow \infty} \frac{f(a + v_j) - f(a)}{v_j}. \end{aligned}$$



Taking  $\xi = \pi_{\mathcal{U}}\langle u_j \rangle \in \mathbb{I}$ , and  $\eta = \pi_{\mathcal{U}}\langle v_j \rangle \in \mathbb{I}$ , we see that: the nonstandard condition of differentiability can not be satisfied simultaneously for  $dx = \xi$  and for  $dx = \eta$ , whatever the choice of  $c$ .  $\square$

**Theorem 12 (Chain-Rule)** *If a function  $f$  is differentiable at  $a$ , and a function  $g$  is differentiable at  $b = f(a)$ , then the composite function  $h := g \circ f$  is differentiable at  $a$ , with derivative  $h'(a) = g'(b) * f'(a)$ .*

*Proof* For any  $\xi \in \mathbb{I}$ ,  $\eta \in \mathbb{I}$ , we have:

$$\begin{aligned} \frac{*f(a + \xi) - f(a)}{\xi} &\approx f'(a); \\ \frac{*g(b + \eta) - g(b)}{\eta} &\approx g'(b). \end{aligned}$$

Let now  $\eta = *f(a + \xi) - f(a)$ . Then:

$$\begin{aligned} \frac{*h(a + \xi) - h(a)}{\xi} &= \frac{*g(*f(a + \xi)) - g(f(a))}{\xi} \\ &= \frac{*g(b + \eta) - g(b)}{\eta} * \frac{\eta}{\xi} \approx g'(b) * f'(a). \end{aligned}$$

Here it is assumed that  $\eta \neq 0$  so as to allow the division by  $\eta$ . But when  $\eta = 0$ , then necessarily  $f'(a) = 0$ , and the proposed identity is trivial.  $\square$

**Theorem 13 (Leibniz' Product Rule)** *If two functions  $f$  and  $g$  are differentiable at  $a$ , then so is their product  $h$ ,  $h(x) = f(x) * g(x)$ . And*

$$h'(a) = f'(a) * g(a) + f(a) * g'(a).$$

*Proof* The non-standard calculations are essentially the same as the standard ones.  $\square$

**Lemma 7** *If  $f$  has a local maximum at  $x = a$ , then*

$$\forall dx \in \mathbb{I}, \quad \Delta_a f(dx) \leq 0.$$

**Theorem 14 ("Fermat")** *On an open interval, if a function  $f$  has a maximum at  $a$ , and there differentiable, then  $f'(a) = 0$ .*

*Proof* Recall the signum function  $\text{sign}(x) = \frac{x}{|x|}$  for  $x \neq 0$ . We will not put the star on its star-extension. For  $dx \in \mathbb{I}$ ,

$$\frac{\Delta_a f(dx) - f'(a)dx}{dx} \in \mathbb{I};$$

So if  $f'(a) \neq 0$ , we take  $dx \in \mathbb{I}$  to have the same signum as  $f'(a)$ , then there is a contradiction:

$$\text{sign}(\Delta_a f(dx)) = \text{sign}(f'(a) * dx) > 0.$$

$\square$

Conventional deductions then give, as corollaries,

**Theorem 15 ("Rolle")** *Let a function  $f$  is continuous on a closed interval  $[a \dots b]$ , and differentiable on the open interval  $(a \dots b)$ . If moreover  $f(a) = f(b)$ , then there exists a zero of the derived function  $f'$  in the open interval. Generally, there exists  $\xi$  in the open interval  $(a \dots b)$ , such that*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

*has a maximum at  $a$ , and there differentiable, then  $f'(a) = 0$ .*

**Remark 5** For a function  $f$  defined on a closed interval, one-sided derivative may still be defined at the endpoints naturally. Nelson mentioned that a stronger notion is in some way more convenient than differentiability.  $f$  is called derivable at a point  $a$ , iff there is a derivative  $f'(a) = c$  such that:

$$(as \ x_1 \rightarrow a, \ x_2 \rightarrow a, \ x_1 \neq x_2),$$

$$\lim \left( \frac{f(x_2) - f(x_1)}{x_2 - x_1} - f'(a) \right) = 0.$$

The NS formulation is:

$$(\epsilon_1 \in \mathbb{I}, \ \epsilon_2 \in \mathbb{I}, \ \epsilon_1 \neq \epsilon_2)$$

$$\Rightarrow \frac{*f(a + \epsilon_2) - *f(a + \epsilon_1)}{\epsilon_2 - \epsilon_1} \approx f'(a).$$

Of course all the differentiation rules are valid for this stronger-sense derivative.

**Theorem 16** *A function  $f$  is derivable on an interval, (i. e., at every point of the interval), iff it is differentiable with continuous derivative everywhere on the interval.*

The function  $f(x) = x^2 \cos(\frac{1}{x})$  (extended by  $f(0) = 0$ ), is differentiable on  $\mathbb{R}$ , but  $f'(x) = 2x \cos(\frac{1}{x}) + \sin(\frac{1}{x})$  (extended by  $f'(0) = 0$ ), is not continuous at  $x = 0$ . Indeed it is not derivable there.

## NS Integral Calculus

Let  $N := (N_j : j \in I) \in \mathbb{N}^I$  be a strictly increasing sequence of natural numbers, and for each  $j \in I$ , define

$$T_j = \left\{ \frac{k}{N_j} : k \in \mathbb{Z}, \ 0 \leq k \leq N_j \right\}.$$

The unit interval  $\mathbb{U}$  is correspondingly partitioned into  $N_j$  subintervals  $[(k-1)/N_j \dots k/N_j]$ ,  $k = 1, 2, \dots, N_j$ . Suppose  $f$  is a function on the unit interval  $\mathbb{U}$ . The left-Riemann sum of  $f$ , relative to this partition is

$$S_j := \frac{1}{N_j} \sum (f(x) : x \in T'_j), \quad T'_j := T_j \setminus \{1\}.$$

Assuming that  $f$  is Riemann-integrable (in the standard sense), this sequence  $S = \langle S_j \rangle$  of Riemann-sums converges to the Riemann integral  $\int_{\mathbb{U}} f(x)dx$ .

**Theorem 17 (Riemann Sum)** *The Riemann-integral for a Riemann-integrable function on the unit interval is the standard part of the “hyperaverage” of  $f$ , over a hyperfinite equi-partition  $\mathcal{T}$  as above:*

$$\int_{\mathbb{U}} f(x)dx = \text{st} \left( \frac{1}{\mathcal{N}} \sum (f(t) : t \in \mathcal{T}) \right).$$

**Remark 6** For the improper Riemann integral  $\int_0^1 (dx)/\sqrt{1-x}$ , the formula is valid.

**Definition 34** Let now  $f: \mathbb{R} \mapsto \mathbb{R}$  be a function, bounded, and compactly supported, which means  $f(x) = 0$  for  $x$  outside a finite interval, then we may define (the “integral over  $\mathbb{R}$ ”)

$$\int f(x)dx := \text{st} \left( \frac{1}{\mathcal{N}} \pi_{\mathbb{U}} \langle S_j \rangle \right);$$

where  $S_j := \sum (f(x) : x \in T_j)$ ;

$$T_j = \left\{ \frac{k}{N_j} : k \in \mathbb{Z}, -N_j^2 \leq k < N_j^2 \right\};$$

And then for any bounded function  $f$  defined on  $\mathbb{R}$ :

$$\int_a^b f(x)dx := \int f(x) * 1_{[a \dots b]}(x)dx.$$

The definition is valid for any function  $f$  as long as it is bounded. (This is just like the Banach-limit for a bounded sequence. Axiom of choice is also involved in the latter!) The Riemann integrability is relevant only to prove that this “integral” is independent of the Riemann-summation scheme. We will now skip all these matter, mentioning only that, as is easily seen:

$$(a < b < c) \Rightarrow \int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx.$$

This justifies the usual definition that

$$\int_a^b f(x)dx := - \int_b^a f(x)dx, \quad \text{for } b < a.$$

### Theorem 18 (Fundamental Theorem of Calculus)

- (1) If  $f$  is continuous on an interval containing a point  $a$ , and  $g(u) := \int_a^u f(x)dx$ , then  $g$  is continuous, and for a continuity point  $b$  of  $f$ , the function  $g$  is differentiable at  $b$ , with  $g'(b) = f(b)$ .
- (2) If  $f$  is derivable on the interval  $[a \dots b]$ , so that the derivative  $f'$  is continuous, hence Riemann integrable, then

$$\int_a^b f'(x)dx = f(b) - f(a).$$

### NS-ODE

**Theorem 19 (Cauchy-Peano)** *If  $f: (x, t) \in \mathbb{R} \times [0 \dots 1] \mapsto f(x, t) \in \mathbb{R}$  is bounded and continuous, then, for any “initial-position”  $a \in \mathbb{R}$ , there is a “solution-path”*

$$y: t \in [0 \dots 1] \mapsto y(t) \in \mathbb{R},$$

such that: it is continuous, and

$$y(0) = a, \quad y'(t) = f(y(t), t), \quad \forall t \in [0 \dots 1].$$

*Proof* The polygonal approximation is the following well-known scheme

$$y = \lim_{j \rightarrow \infty} y_j:$$

Take a natural number  $N_j$ , and divide the time-interval into  $N_j$  equal subintervals. In the duration of  $k$ th subinterval  $(k-1)/N_j \leq t < k/N_j$ , the velocity is approximated by the velocity-field at the latest position:  $y_j(t) := f(y((k-1)/N_j), (k-1)/N_j)$ , and consequently there is a recursion:

$$y_j \left( \frac{k}{N_j} \right) := y_j \left( \frac{k-1}{N_j} \right) + \frac{1}{N_j} * f \left( y_j \left( \frac{k-1}{N_j} \right), \frac{k-1}{N_j} \right); \quad k = 1, 2, \dots, N_j.$$

The standard way is to take a “limiting point” of the sequence of approximate solutions (which are polygonal = “piecewise-linear”),  $y_j(t)$ ,  $j = 1, 2, 3, \dots$ . They are considered as points of the path-space (= function space, with uniform topology). Ascoli’s compactness criterion ensures the existence of such a limiting point  $y_\infty$ . From the uniform convergence

$$\lim_{j \rightarrow \infty} y_j(t) = y_\infty(t),$$

$$\lim_{j \rightarrow \infty} f(y_j(t), t) = f(y_\infty(t), t),$$

it follows easily the sought-for identity:

$$y_{\infty}(t) = a + \int_0^t f(y_{\infty}(s), s) ds.$$

The standard proof is translated into the NS proof thus: With  $N_j \uparrow \infty$ , we have an internal hyperfinite number (unlimited, of course),  $\mathcal{N} = \pi_{\mathcal{U}}(N)$ ,  $N := \langle N_j \rangle$ , together with the internal set of partition-points on the NS time-interval:

$$\mathcal{T} := \left\{ \frac{k}{\mathcal{N}} : k \in \mathbb{Z}, 0 \leq k \leq \mathcal{N} \right\}.$$

For “time”  $\tau = \frac{k}{\mathcal{N}}$  on this hyperfinite set, define the  $\mathcal{N}$ th approximate “position”  $Y(\tau)$  to be

$$Y\left(\frac{k}{\mathcal{N}}\right) = a + \frac{1}{\mathcal{N}} * \sum_{i=0}^{k-1} f\left(Y\left(\frac{i}{\mathcal{N}}\right), \frac{i}{\mathcal{N}}\right).$$

Finally take

$$y(t) := \text{st} \left( Y\left(\frac{\text{floor}(t * \mathcal{N})}{\mathcal{N}}\right) \right).$$

The above NS summation-multiplied-by-infinitesimal goes to Riemann integral:

$$y(t) = a + \int_0^t f(y(s), s) ds.$$

□

It is well-known that under the very general condition of boundedness and continuity, the solution is not unique, reflecting the arbitrariness in the choice of one limiting point from a compact set. Of course this corresponds to the arbitrariness of the choice of the unlimited  $\mathcal{N} \in {}^*\mathbb{N}$ , or rather, the hyper-finite partition  $\mathcal{T}$ .

## General Idea of NSA

### The Set-Terminology

Ever since G. Cantor invented the Mengenlehre, and after the Hilbertian axiomatization (r) evolution, the naive set-theory became the common language of mathematics. Everything is (to be expressed in terms of) set! As an example, the color “blue” may be defined as “the set of all blue things”. (I’d like to make this as my first teaching of pedagogy.) If this sentence is hardly comprehensible, then an easier explanation is as follows. Suppose I say: “This stone  $\omega$  is blue”. Then what I mean is the mathematical sentence “ $\omega \in B$ ”, where  $B$  is a conveniently collected set of blue

stones. Of course,  $B$  should be a subset of a convenient “total set”  $\Omega$  of stones. With this primitive idea of ‘membership relation’, simple set-notions fit in with our daily logic. Precisely: if  $a$  is the sentence “ $x \in A$ ”, and  $b$  is “ $x \in B$ ”, then the logical combinations of assertions correspond to the Boolean combinations of sets in the well-known way:

“ $a$  or  $b$ ” is “ $x \in A \cup B$ ”;

“ $a$  and  $b$ ” is “ $x \in A \cap B$ ”;

“not  $b$ ” is “ $x \in \Omega \setminus B$ ”.

Here a convenient total set  $\Omega$  should be available for the logical analysis. From this thinking, a set  $\mathcal{A}$  of subsets of  $\Omega$  is called a Boolean algebra of sets, if it is closed under the three set-operations mentioned above:  $\cup$ ,  $\cap$ , and “complementation” (set-difference by the total set). My second example of the “triumph of naive set theory” is the axiomatization of probability theory by A. N. Kolmogorov. Its publication was 10 years after N. Wiener’s construction of Wiener process. Wiener wanted to talk about the “probabilistic behavior of the path  $W$  of a particle”. By a path is meant a function from the time-axis to the state-space,  $W(t)$  signifying “the position at time  $t$ ”. Here the time-axis is the half-line  $[0 \cdots + \infty)$ , and the state-space is the real line  $\mathbb{R}$ . And Wiener requires that the path is continuous, and starts from the origin:  $X = 0$ . In the view point of Kolmogorov, the phrase “probabilistic behavior of the path  $W$  of a particle” should be changed to “behavior of a random particle”. “A random particle” just means that it is “one chosen from many”; these many form a set. (This set, the so-called probability space, is always an imagined set. It is almost always an infinite set, so never “concrete” for students of probability theory!) The picture is: Lady Luck collects a big set  $\Omega$  of all particles (= chance-lings). For each particle  $\omega \in \Omega$ , its path is  $W(\cdot, \omega)$ , mapping time  $t$  to its position  $W(t, \omega)$  at time  $t$ . (So Wiener process  $W$  is a “two-variable function”, its domain of definition is the set-product  $\mathbb{R}_+ \times \Omega$ . Textbooks of probability theory always suppress the variable  $\omega$  which signifies the randomness.) The success of Kolmogorov’s axiomatization is simply this: it makes precise what kind of “questions about probability” one can ask. Now probability comes in, only because Lady Luck chooses one chance-ling  $\omega$  from all conceivables. (Then this chosen one is the “reality”). We are interested (only) in the assertions of this form “ $\omega \in B$ ”. We can only ask questions of this kind: “What is the probability of the event  $\omega \in B$ ?” And that only for some kind of set  $B$ . The answer  $\mu(B)$  is the probability predestined by Her. For what kind of  $B \subset \Omega$  is  $\mu(B)$  meaningful is also Her designation. Then we simply call  $B$  an event. So  $\mu(B)$  is now the probability of event  $B$ . (In-

stead of the “event  $\omega \in B$ ”). Since we can do Boolean combinations of assertions, correspondingly, the set  $\mathcal{A}$  of all events should be a Boolean set-algebra. But in probability theory, always the stronger condition of sigma-completeness is required:  $\mathcal{A}$  is closed under countable union, thus called a  $\sigma$ -algebra.

**Definition 35 ((Kolmogorov) Probability Space)**

A probability space is a nonempty set  $\Omega$ , together with a measure structure  $\langle \mathcal{A}, \mu \rangle$ , where  $\mathcal{A}$  is a  $\sigma$ -algebra on  $\Omega$ , and  $\mu: \mathcal{A} \mapsto \mathbb{U}$  satisfies the countable-additivity condition: If  $\forall n, B_n \in \mathcal{A}$  and  $\forall m \neq n, B_m \cap B_n = \emptyset$ , then

$$\mu(\cup_{1 \leq n < \infty} B_n) = \sum_{n=1}^{\infty} \mu(B_n).$$

Returning to Wiener’s process. Lady Luck should allow assertions  $\omega \in B$  of the form  $W(t, \omega) \in (c_1 \dots c_2]$ , where time  $t \in [0 \dots + \infty)$  and interval  $(c_1 \dots c_2]$  are arbitrary. With this condition satisfied, the two-variable function  $W: \langle t, \omega \rangle \mapsto W(t, \omega)$  is called a stochastic process on the probability space  $(\Omega, \mathcal{A}, \mu)$ . Actually the central questions asked by (Bachelier and) Wiener are events of the following forms:

$$B := \{\omega \in \Omega: \text{ for } k = 1, 2, \dots, n, W(t_k, \omega) - W(s_k, \omega) \in A_k\},$$

where

$$s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_n < t_n; \\ A_1, A_2, \dots, A_n, \text{ are intervals}$$

**Theorem 20 (Wiener)** *There exists a probability space  $(\Omega, \mathcal{A}, \mu)$ , and a stochastic process  $W$  on it, such that: for each  $\omega \in \Omega$ , the sample-path  $W(\cdot, \omega)$  is a continuous function on  $[0 \dots + \infty)$ ; and, for set  $B$  as above,*

$$\mu(B) = \prod_{1 \leq k \leq n} \frac{1}{\sqrt{2\pi(t_k - s_k)}} \int_{A_k} e^{\frac{-x^2}{2(t_k - s_k)}} dx.$$

**Naive Set Operations**

As we have explained above, a mathematical system is a set with structure, which is a finite set of items of functions, relations, or functions. We now explain that starting from a ground set, there is a super-structure, which will contain all conceivable structures. Naive as we are, sometimes we just have to be very careful. So we give a very careful definition of couple and relation, which will be a foundation of all that follows.

**Definition 36 (Couple and Cartesian Product)** For two non-empty sets  $A, B$ , the Cartesian product is defined as

$$A \times B := \{\langle a, b \rangle: a \in A, \text{ and } b \in B\}.$$

Here “a couple”, or “an ordered pair”, is defined by

$$\langle a, b \rangle := \{\{a\}, \{a, b\}\}.$$

The ordered  $n$ -tuple, for  $n > 2$ , may be defined recursively as

$$\langle x_1, x_2, \dots, x_n \rangle := \langle \langle x_1, x_2, \dots, x_{n-1} \rangle, x_n \rangle.$$

Note that we could write the “1-tuple” as  $\langle a \rangle := a$ . Then write:

$$A^n := \{\langle x_1, x_2, \dots, x_n \rangle: x_1, x_2, \dots, x_n \in A\}.$$

Thus  $A \times B \times C$  is defined to be  $(A \times B) \times C$ . (And so forth!)

The key to this definition is the observation that

$$\langle \langle x, y \rangle = \langle u, v \rangle \rangle \text{ iff } ((x = u) \text{ and } (y = v)).$$

Here it is to be emphasized that when forming a set, multiple enumeration of any element just counts once, but does no harm. And in the case of forming the pair  $\langle a, b \rangle$ , when  $a = b$ , the pair is reduced to  $\langle a, a \rangle := \{\{a\}\}$ , and this is a singleton set whose only element is the singleton set  $\{a\}$ .

**Definition 37 (Relation)** If  $R \subset A \times B$ , then  $R$  is a relation from  $A$  to  $B$ . And in case  $A = B$ , it is called a relation on  $A$ . The domain of  $R$  is the set

$$\text{Dom}(R) := \{x \in A: \exists y \in B, \langle x, y \rangle \in R\}.$$

If  $\langle x, y \rangle \in R$ , sometimes we write (the “infix notation”!)  $xRy$ , still sometimes (the “prefix notation”!)  $R(x, y)$ . Suppose  $R$  and  $S$  are two relations, and  $R \subset S$ , then  $R$  is a restriction of  $S$ , and  $S$  an extension of  $R$ .

**Definition 38 (Function and Strict-Evaluation)** If  $f$  is a relation from  $A$  to  $B$ , and, for any  $x \in A$ , there is exactly one  $y \in B$ , such that  $\langle x, y \rangle \in f$ , then we call  $f$  a function. For a function  $f$  and an element  $x$  in its domain of definition, the ‘evaluation’ gives the value  $f(x)$  of  $x$  under  $f$ . So this may be generalized in the following way: if  $s, R$  are fixed, and there is exactly one  $t$ , such that:  $\langle s, t \rangle \in R$ , we then write:

$$t := R \uparrow s.$$

**Definition 39 ( $n$ -ary Relation)** If  $R \subset A^n$ ,  $R$  is an  $n$ -ary relation on  $A$ . How to define a two-variable function? Let

us take the arithmetic functions  $+$  (addition), and  $*$  (multiplication), as examples. They are from  $\mathbb{R} \times \mathbb{R}$  to  $\mathbb{R}$ , are usually infix-written, thus

$x + y$  is precisely  $\text{add } \vec{\uparrow} \langle x, y \rangle$ ;

$x * y$  is precisely  $\text{multiply } \vec{\uparrow} \langle x, y \rangle$ .

**Definition 40 (Power Set, Exponential Set)** The set of all subsets of a set  $A$ , (including the empty set  $\emptyset$ , and the total set  $A$ ), is denoted by  $\mathbb{P}(A)$ , and called the power-set of  $A$ . Starting with  $\mathbb{P}^1(A) := \mathbb{P}(A)$ , iteration of this power-set construction gives

$$\mathbb{P}^{n+1}(A) = \mathbb{P}(\mathbb{P}^n(A)), \quad \text{and we add } \mathbb{P}^0(A) = A.$$

Consider the set  $A = \mathbb{D}_n := \{0, 1, \dots, n-1\}$ , which has cardinality  $n \in \mathbb{N}$ , then  $\mathbb{P}(A)$  has cardinality  $2^n$ . We think “power-set” is an unfortunate terminology, in our opinion. A better one is the “exponential set”, and a better notation is  $2^A$ . 2 is, to be sure, the natural base for the discrete mathematics. Or, if (2 being identified with the set  $\mathbb{D}_2$  above),  $f \in 2^A$  is interpreted as a function from  $A$  to  $\mathbb{D}_2$ , it is surely the indicator function of a subset of  $A$ . On the other hand, sets like  $A^2, A^3, \dots$ , should be called the power sets of  $A$ . ( $A$  to the 2nd, 3rd, power,  $\dots$ ) It is generally known that in naive set theory, it is illegal to talk about “the set of all sets”. A convenient rule is that any set cannot be an element of itself:  $A \notin A$ . For any set  $A$ , then  $A \neq \{A\}$ , and indeed the latter is a singleton set. Now there is no entity as a set containing everything, there are sometimes some entity which can not be a set.

**Definition 41 (Ur-element)** Sometimes a non-empty set  $S$  may be declared as a ground-set, and its elements declared as ur-elements. A mathematical entity  $x$  is declared as an ur-element (or individual), if a sentence  $y \in x$  is never allowed true. (A mathematical entity is either an ur-element or a set).

An ur-element is just like a sodium ion which is declared decomposable in high-school or freshman Chemistry. This is the most convenient concept for the students.

**Definition 42 (Super-structure)** Suppose  $S$  is a set of ur-elements. We define  $S_0 := S$ , and then recursively:

$$S_n := S_{n-1} \cup \mathbb{P}(S_{n-1}).$$

The set

$$S_\infty := \bigcup_{j \geq 0} S_j$$

is called the super-structure over the ground-set  $S$ . Any  $x \in S_\infty$  is assigned a “rank”

$$\min\{j: x \in S_j\}.$$

This rank is positive, iff  $x$  is a set.

**Notation 2** The following are the symbols for logical constructions:

- The symbol  $\neg$  is read as “the negation of”.
- The symbol  $\wedge$  is read as “and”.
- The symbol  $\vee$  is read as “or”.
- The symbol  $\Rightarrow$  is read as “imply”.
- The symbol  $\exists$  is read as “there exists”.
- The symbol  $\forall$  is read as “for all”.

**Theorem 21** Every concept in mathematics can be explained in terms of set. In mathematical discussion, we take a suitable ground set  $S$ , build up the super-structure  $S_\infty$ , then the simplest kinds of assertions are: “ $a \in b$ ”, or “ $a = b$ ”, where  $a$  and  $b$  are elements of  $S_\infty$ . From these simple sentences more complicated sentences are formed with the aid of logical symbols.

**Example 7 (Group)** Consider a quadruple  $\langle G, e, \Lambda, * \rangle$ , where  $G$  is a set,  $e, \Lambda, *$  are respectively 0-, 1-, 2-variable functions with a triple  $\langle e, \Lambda, * \rangle$ , on  $G$ . The group axiom for this system is the sentence:

$$\begin{aligned} & (\forall x \in G)(\forall y \in G)(\forall z \in G)((x * (y * z) \\ & = ((x * y) * z) \wedge ((e * x) = x) \wedge (\hat{x} * x = e)). \end{aligned}$$

This is too formal. But I hope by now you will have no real obstacle in understanding and accepting it. We see that if  $G$  is contained in the ground-set  $S_0$ , then  $e, \Lambda, *$  are all in the superstructure  $S_\infty$  with ranks 0, 3, 5, respectively. It is important to realize that group theory should be defined as the study of the true-or-false of those sentences involving only the group structure. Suppose  $\Phi: G \mapsto H$  is group-homomorphism. Then a sentence (about group) true in  $G$  will be true in  $\Phi[G] \subset H$ , after the sentence is transferred (or translated, or reinterpreted), by  $\Phi$ . Commutativity is an instance:

$$(\forall x \in G)(\forall y \in G), \quad (x * y = y * x).$$

It is possible that  $G$  is commutative, and  $H$  is not, but the homomorphic image  $\Phi[G](\subsetneq H)$  is commutative. Any sentence (about group) valid in  $G$  is valid in  $\Phi(G)$ .

**Theorem 22 (NSA Idea)** The star-extension from the superstructure to the hyper-universe is a “homomorphism”.

The key ideas of NSA, in the superstructure approach is the following:



- From a ground set  $S_0$  of ur-elements, we can build up set-theoretically the super-structure

$$S_\infty := \bigcup_{j \geq 0} S_j,$$

the “standard universe”. All the mathematical entities arising from  $S_0$  are in it.

- Associated with  $S_0$ , there is the corresponding non-standard ground-set  $^*S_0 = S_0^{\mathfrak{U}}$  of ur-elements. But

$$S_\infty^{\mathfrak{U}} := \bigcup_{n \geq 0} S_n^{\mathfrak{U}},$$

the superstructure of  $S_0^{\mathfrak{U}}$ , is not interesting, because most of the entities in it are not very interesting: they are external to us. Instead of it, and inside of it, the hyper-universe (or non-standard universe)

$$^*S_\infty := \bigcup_{n \geq 0} ^*S_n,$$

of internal entities, is of central importance in NSA. Entities in  $S_\infty^{\mathfrak{U}} \setminus ^*S_\infty$  are then called external.

- There is an injective mapping from  $S_\infty$  into  $^*S_\infty$ , called the star-extension. Every standard entity  $A \in S_\infty$  is star-mapped to its star-extension  $^*A \in ^*S$ , which is its (non-standard or) hyper version, or star version in the hyper-universe. All these are the (hyper versions of) standard entities. Other entities of the hyper-universe  $^*S_\infty$  are internal but non-standard.
- There is a Transfer Principle, which says that the star-embedding is the all-encompassing “injective homomorphism”: If a sentence is true in the standard interpretation in terms of the standard universe, then this sentence holds true in the non-standard interpretation in terms of the hyper-universe. By “non-standard interpretation” we mean that all the constant entities occurring in the sentence are to be replaced by their star-extensions. (So a quantifying clause “ $\forall x \in A$ ” is replaced by “ $\forall x \in ^*A$ ”).

### The Ultra-Power Construction

*Note 3* The notation  $S_0, S_n, S_\infty$  have the meaning as defined above. (So  $S_0$  is a ground-set of ur-elements). There will be only another set  $S_0^{\mathfrak{U}}$  of ur-elements! Any appearance like  $Z_n, ^*S_n$ , shall not be interpreted as part of super-structure construction!  $\mathfrak{U}$  is a free ultra-filter on a set  $I$  of indices, to which all the phrases “ultimate” refer.

- Let  $Z_n$  be the set of “sequences”  $f = \langle f_j \rangle$  such that for  $j, f_j \in S_n$ . The sequence  $Z_n$  is increasing with  $n$ , with union

$$Z_\infty = \bigcup_{n \geq 0} Z_n.$$

- For each  $n \in \mathbb{N}_0$ , on the set  $Z_n$ , we define an equivalence-relation  $\sim^{\mathfrak{U}}$  (as before), by:  $(f \sim^{\mathfrak{U}} g)$  iff:

$$f_j = g_j \quad \text{for ultimate } j.$$

Obviously we see that attaching a suffix  $n$  to the equivalence relation is hardly necessary:  $\sim^{\mathfrak{U}}$  for  $Z_{n+1}$  is an extension of  $\sim^{\mathfrak{U}}$  for  $Z_n$ .

- So we may define  $S_0^{\mathfrak{U}} = \pi_{\mathfrak{U}}(Z_0)$  to be the quotient set of  $Z_0$  under this equivalence relation  $\sim^{\mathfrak{U}}$ , i. e., the set of all equivalence-classes of  $Z_0$ . Now we declare that  $S_0^{\mathfrak{U}}$  be a ground-set of ur-elements, so that a formula like  $y \in x$  is for  $x \in S_0^{\mathfrak{U}}$  is illegal. We then build up the superstructure over  $S_0^{\mathfrak{U}}$  as before:

$$S_\infty^{\mathfrak{U}} := \bigcup_{n \geq 0} S_n^{\mathfrak{U}}; \quad S_{n+1}^{\mathfrak{U}} := S_n^{\mathfrak{U}} \cup \mathbb{P}(S_n^{\mathfrak{U}}).$$

- The canonical projection from  $Z_0$  into  $S_0^{\mathfrak{U}}$  (surjectively) being here denoted by  $\pi_{\mathfrak{U}}$ , we now recursively extend it to a mapping

$$\pi_{\mathfrak{U}}: f \in Z_n \mapsto \pi_{\mathfrak{U}}(f) \in S_n^{\mathfrak{U}}, \quad \forall n \in \mathbb{N}.$$

After  $\pi_{\mathfrak{U}}(g) \in S_{n-1}^{\mathfrak{U}}$  are defined for every  $g \in Z_{n-1}$ , we then define  $\pi_{\mathfrak{U}}(f)$  for  $f \in Z_n \setminus Z_{n-1}$  to be

$$\pi_{\mathfrak{U}}(f) := \{ \pi_{\mathfrak{U}}(g) : g \in Z_{n-1}, \text{ and } g_j \in f_j, \text{ for ultimate } j \}.$$

It is easy to see that  $\pi_{\mathfrak{U}}$  is well-defined on  $S_\infty^{\mathfrak{U}}$ , and, by this definition, all over each  $S_n^{\mathfrak{U}}$ , and thus over  $S_\infty^{\mathfrak{U}}$ , we have

$$(\pi_{\mathfrak{U}}(f) = \pi_{\mathfrak{U}}(h)) \iff (f \sim^{\mathfrak{U}} h).$$

The proof of this (and similar ones below) is tedious but easy consequence of the dichotomy condition of the ultra-filter.

- Finally we have the hyper-universe of internal entities:

$$^*S_\infty := \bigcup_{n \geq 0} ^*S_n; \quad ^*S_n := \{ \pi_{\mathfrak{U}}(f) : f \in Z_n \} \subset S_n^{\mathfrak{U}}.$$

Any set  $B \in S_\infty^{\mathfrak{U}} \setminus ^*S_\infty$  is called strictly external.

- And we see that (just as what we have done before for  $I = \mathbb{N}$ ):

$$\begin{aligned} \pi_{\mathfrak{U}}\langle A_j \rangle \cap \pi_{\mathfrak{U}}\langle B_j \rangle &= \pi_{\mathfrak{U}}\langle A_j \cap B_j \rangle; \\ \pi_{\mathfrak{U}}\langle A_j \rangle \cup \pi_{\mathfrak{U}}\langle B_j \rangle &= \pi_{\mathfrak{U}}\langle A_j \cup B_j \rangle; \\ \pi_{\mathfrak{U}}\langle A_j \rangle \setminus \pi_{\mathfrak{U}}\langle B_j \rangle &= \pi_{\mathfrak{U}}\langle A_j \setminus B_j \rangle; \end{aligned}$$

And the set of all internal sets form a Boolean set-ring.

- For  $\{f, g, h\} \subset Z_\infty$ ,

$$\pi_{\mathcal{U}}(h) = \{\pi_{\mathcal{U}}(f), \pi_{\mathcal{U}}(g)\} \quad \text{iff } h_j = \{f_j, g_j\},$$

for ultimate  $j$ ;

$$\pi_{\mathcal{U}}(h) = \langle \pi_{\mathcal{U}}(f), \pi_{\mathcal{U}}(g) \rangle \quad \text{iff } h_j = \langle f_j, g_j \rangle,$$

for ultimate  $j$ ;

$$\pi_{\mathcal{U}}(h) = \pi_{\mathcal{U}}(f) \dot{\vdash} \pi_{\mathcal{U}}(g) \quad \text{iff } h_j = f_j \dot{\vdash} g_j,$$

for ultimate  $j$ .

- For any  $n \in \mathbb{N}_0$ , if  $A \in S_n$ , it may be identified with the “sequence” with constant term  $A$ , i. e., we may consider  $S_n \subset Z_n$ . Then  $\pi_{\mathcal{U}}(A) \in {}^*S_n$  is written as  ${}^*A$ , and called a standard internal entity, or a star version of  $A$ , in the hyper-universe, as was mentioned before. Also note that: though  $S_\infty \not\subset {}^*S_\infty$ , and  ${}^*S_\infty \not\subset {}^*S_\infty$ , the star notation is still compatible.

**Remark 7** If  $b \in A \in S_\infty$ , then the standard entity  ${}^*b \in {}^*A$ . That is

$${}^\sigma A := \{{}^*b : b \in A\} \subset {}^*A.$$

All standard members in  ${}^*A$  are gathered into the set  ${}^\sigma A$ . Indeed when  $A$  is a finite set,  ${}^*A = {}^\sigma A$ . But when  $A$  is an infinite set, the right side above is properly contained in the left-side set. That is: the set  ${}^*A$  must contain non-standard elements. Then both  ${}^\sigma A$  and  ${}^*A \setminus {}^\sigma A$  are external sets.

This is the source of all confusions!

- Internality is transitive in the sense that

$$(x \in y \quad \text{and} \quad y \in {}^*S_\infty) \Rightarrow (x \in {}^*S_\infty).$$

- For entities  $A, B, C$  in  $S_\infty$ ,

$$A = B \quad \text{iff} \quad {}^*A = {}^*B;$$

$$A \in B \quad \text{iff} \quad {}^*A \in {}^*B;$$

$$C = \langle A, B \rangle \quad \text{iff} \quad {}^*C = \langle {}^*A, {}^*B \rangle;$$

$$C = A \dot{\vdash} B \quad \text{iff} \quad {}^*C = {}^*A \dot{\vdash} {}^*B;$$

For sets  $A, B$ , in  $S_\infty \setminus S_0$ ,

$${}^*(A \cup B) = {}^*A \cup {}^*B;$$

$${}^*(A \cap B) = {}^*A \cap {}^*B;$$

$${}^*(A \setminus B) = {}^*A \setminus {}^*B;$$

$${}^*(A \times B) = {}^*A \times {}^*B;$$

**Theorem 23 (Transfer Principle)** *If a sentence is true in the standard universe  $S_\infty$ , then its star-version is true in the hyper-universe  ${}^*S_\infty$ .*

All simple mathematical sentences are assertions about either the identity of two mathematical entities, or the membership of one against the other. By the use of “logical connectives”, more complicated assertions are then constructed. So the principle means exactly the whole set of “homomorphisms” displayed above.

**Theorem 24 (Standardization Principle)** *Let  $A \in S_\infty$  be a standard entity, and  ${}^*A \in {}^*S_\infty$  its star-version. Let  $B \in S_\infty^{\mathcal{U}} \setminus S_0^{\mathcal{U}}$  be any non-standard set. Then there is  $C \in S_\infty$ , such that: for any  $x \in S_\infty$ ,*

$$({}^*x \in {}^*C) \quad \text{iff} \quad ({}^*x \in ({}^*A \cap B)).$$

*$C$  is called the standard part of  $A$ . The proof is simple:  $C$  is the union of all subsets  $D \subset A$ , such that*

$$({}^*x \in {}^*D) \Rightarrow ({}^*x \in ({}^*A \cap B)).$$

**Definition 43 (Concurrence Relation)** A relation  $R \in S_\infty \setminus S_1$  is called concurrent, if: for any finite number of elements  $a_1, a_2, \dots, a_n$  in the domain of  $R$ , there is one  $b \in S_\infty$ , such that

$$\langle a_i, b \rangle \in R, \quad \text{for} \quad i = 1, 2, \dots, n.$$

**Example 8** An obvious example of concurrence relation is the “less than” relation  $<$  in the set  $\mathbb{N}$ . In this case, there is an unlimited hyperinteger  $\omega$  bigger than all standard integer  $a \in \mathbb{N}$ , that is:  $\forall a \in \mathbb{N}, a < \omega$ . All we need, in this particular example, is that the indicial set  $I$  has infinite cardinal.

More generally, we need an index set  $I$  with big enough cardinal. Then the following Concurrence Theorem is valid. (But the proof is bigger than our concern).

**Theorem 25 (Idealization Principle)** *Let  $R \in S_\infty$  be a concurrent relation. Then there is  $b \in {}^*S_\infty$ , such that for all  $a \in \text{Dom}(R)$ ,  $\langle a, b \rangle \in R$ .*

**Note 4 (Internal Set Theory)** There are now some axiomatic approach of NSA. In the internal set theory proposed by E. Nelson, the notion of “standardness” is considered as fundamental and undefined, and the three principles of idealization, standardization, and transfer, are taken as axioms, (in addition to the Fraenkel–Zermelo axioms with Axiom of Choice), for the NSA. This is very elegant, though not our choice in this article.

### Some Consequences of Transference

**Example 9 (Well-ordering in  ${}^*\mathbb{N}$ )** The axiom of mathematical induction for  $\mathbb{N}$  is equivalent to the well-ordering

property of  $\mathbb{N}$ :

$$((\forall A \subset \mathbb{N}) \wedge (A \neq \emptyset)) \Rightarrow (\exists m \in A, (\forall n \in A, m \leq n)).$$

“Every non-empty subset  $A$  of  $\mathbb{N}$  has a minimum element”.

The transferred sentence is not:

$$((\forall A \subset {}^*\mathbb{N}) \wedge (A \neq \emptyset)) \Rightarrow (\exists m \in A, (\forall n \in A, m \leq n)).$$

“Every non-empty subset  $A$  of  ${}^*\mathbb{N}$  has a minimum element”. Instead, the correct “translation” of this sentence is “Every non-empty internal subset  $A$  of  ${}^*\mathbb{N}$  has a minimum element”. Why? We may write the original premise as  $\forall A \in \mathcal{P}(\mathbb{N})$ , then, in the transference, not only the set  $\mathbb{N}$  is star-extended to  ${}^*\mathbb{N}$ , but also its power-set is star-extended to the standard set (in non-standard version),  ${}^*\mathcal{P}(\mathbb{N})$ .

$$\begin{aligned} ((\forall A \subset {}^*\mathcal{P}(\mathbb{N})) \wedge (A \neq \emptyset)) \\ \Rightarrow (\exists m \in A, (\forall n \in A, m \leq n)). \end{aligned}$$

Another way of thinking is to star-transform the original premise “ $\forall A \subset \mathbb{N}$ ” into “ $\forall A \subset {}^*\mathbb{N}$ ”. (That is: the subset-relation gets star-extended.) In the ultra-power representation used before,  $A = \pi_{\mathcal{U}} \langle A_j \rangle$ , then take  $m_j = \min(A_j) \in \mathbb{N}$ , which exists because of the well-ordering of  $\mathbb{N}$ , and  $\mathbb{N} \supset A_j \neq \emptyset$ . Then  $m := \pi_{\mathcal{U}} \langle m_j \rangle \in {}^*\mathbb{N}$  is the minimum of  $A$  sought for. The wrong transference gives us the well-ordering property of  ${}^*\mathbb{N}$ , which is obviously wrong, because the set of all unlimited hyper natural numbers  ${}^*\mathbb{N} \setminus \mathbb{N}$  would have a minimum element  $\zeta$ , which is absurd because  $\zeta - 1$  would have been still an element of this set.

**Corollary 2** *The set  $\mathbb{N}$  is an external subset of  ${}^*\mathbb{N}$ . So is the complement  ${}^*\mathbb{N} \setminus \mathbb{N}$ .*

**Example 10 (Order-completeness of  ${}^*\mathbb{R}$ )** The order-completeness of  $\mathbb{R}$  says that: “Every non-empty subset  $A$  of  $\mathbb{R}$  which is bounded above has a supremum (= sup = least upper bound)”.

$$\begin{aligned} ((\forall A \subset \mathbb{R}) \wedge (A \neq \emptyset) \wedge (\exists b \in \mathbb{R}, (\forall x \in A, x \leq b)) \\ \Rightarrow ((\exists m \in \mathbb{R}, (\forall x \in A, x \leq b)) \\ \wedge (\forall r \in \mathbb{R}, ((\forall x \in A, x \leq r) \Rightarrow m \leq r))). \end{aligned}$$

The transferred sentence is not the order-completeness of the field  ${}^*\mathbb{R}$ : “Every non-empty subset  $A$  of  ${}^*\mathbb{R}$  which is bounded above has a supremum”. A ready counterexample is the set of infinitesimals  $\mathbb{I}$ .  $\mathbb{I}$  is certainly bounded above by any strictly positive standard real number. But it

does not have a “supremum  $\zeta$ ”. Otherwise,  $\zeta$  could only be positive. If it is a positive real number, then  $\frac{\zeta}{2}$  is a smaller upper-bound. If it is a positive infinitesimal, then an infinitesimal, namely  $2 * \zeta$  could not be bounded by  $\zeta$ . (We may also appeal to the theorem that there is only one order-complete field, up to a canonical isomorphism.) Again the correct transference should replace “subset” by “internal subset”: “In  ${}^*\mathbb{R}$ , every non-empty internal subset  $A$  which is bounded above has a sup (= least upper bound)”.

**Corollary 3 (Overflow and Underflow)** *Let  $A$  be an internal subset of  ${}^*\mathbb{R}$ .*

- (1): *If for any  $n \in \mathbb{N}$ , there is a limited  $x \in A$ , such that  $x > n$ , then there is unlimited positive  $\eta \in A$ .*
- (2): *If for any unlimited positive  $\zeta \in A$ , there is another unlimited  $\xi \in A$ , such that  $\xi < \zeta$ , then there is a limited positive  $\eta \in A$ .*
- (3): *Both subsets  $\mathbb{R} \subset {}^*\mathbb{R}$ , and  ${}^*\mathbb{R} \setminus \mathbb{R}$ , of the standard set  ${}^*\mathbb{R}$  (hyper-version) are external.*

*Proof*

- (1) We may assume that  $A$  is bounded above. But then with  $\xi = \sup(A) \in {}^*\mathbb{R}$ ,  $\xi$  is unlimited. Therefore by the definition of lub, there exists an  $\eta \in A$ , with  $\frac{\xi}{2} \leq \eta \leq \xi$ .
- (2) Let  $z$  be the infimum (= inf = greatest lower bound) of the subset of positive elements of  $A$ , then  $z$  is limited. And by the definition of inf, there is  $\eta \in A$ , such that  $z \leq \eta \leq z + 1$ .  $\square$

**Theorem 26 (Saturation Principle)** *Let  $\langle A_n : n \in \mathbb{N} \rangle$  be a sequence of internal sets with the “finitely-intersecting property”:*

$$\forall n \in \mathbb{N}, \quad \bigcap_{1 \leq m \leq n} A_m \neq \emptyset.$$

*Then the sequence is itself intersecting:*

$$\bigcap_{m \in \mathbb{N}} A_m \neq \emptyset.$$

(For the proof, the countable indicial set  $I = \mathbb{N}$  is enough.) Now for each  $n \in \mathbb{N}$ ,  $A_n = \pi_{\mathcal{U}} \langle A_{n,j} : j \in I \rangle$ . And for this fixed  $n \in \mathbb{N}$ ,

$$\bigcap_{m \in \mathbb{N}} A_m = \pi_{\mathcal{U}} \langle \bigcap_{1 \leq m \leq n} A_{m,j} : j \in I \rangle \neq \emptyset \in {}^*S_{\infty}.$$

Therefore for each  $n \in \mathbb{N}$ , and for ultimate  $j$ ,  $\bigcap_{1 \leq m \leq n} A_{m,j} \neq \emptyset$ . (In particular, we may and will just assume: for all  $j$ ,  $A_{1,j} \neq \emptyset$ ). For each  $j \in I$ , we define

$$k_j := \max\{n \in \mathbb{N} : \bigcap_{1 \leq m \leq n} A_{m,j} \neq \emptyset, \text{ and } n \leq j\};$$

Now for each  $j \in I$ , choose an  $x_j \in \cap_{1 \leq m \leq k_j} A_{m,j}$ . We see that  $\xi = \pi_{\mathcal{U}} \langle x_j \rangle \in A_n$ , for any  $n \in \mathbb{N}$ . Indeed for any  $n \in \mathbb{N}$ ,

$$\begin{aligned} & \{j \in I: x_j \in A_{n,j}\} \supset \{j: k_j \geq n\} \\ & = \{j \in I: j \geq n\} \cap \{j \in I: \cap_{m \leq n} A_{m,j} \neq \emptyset\} \in \mathcal{U}. \end{aligned}$$

**Corollary 4** For a sequence  $\langle A_n: n \in \mathbb{N} \rangle$  of internal sets, their (countably infinite) union is internal  $\cup_{n \in \mathbb{N}} A_n \in {}^*S_\infty$ , iff the union equals to a finite sub-union. I. e., iff there exists  $m \in \mathbb{N}$ , such that

$$\cup_{n \in \mathbb{N}} A_n = \cup_{1 \leq n \leq m} A_n.$$

*Proof* If  $B := \cup_{n \in \mathbb{N}} A_n \in {}^*S_\infty$  (is internal), then: all  $B_n := B \setminus A_n$  are internal, and  $\cap_{n \in \mathbb{N}} B_n = \emptyset$ , by de Morgan's law. Then the saturation principle requires that for a certain  $m \in \mathbb{N}$ ,  $\cap_{1 \leq n \leq m} B_n = \emptyset$ . Or,  $\cap_{1 \leq n \leq m} (B \setminus A_n) = \emptyset$ , and  $\cup_{1 \leq n \leq m} A_n = B = \cup_{n \in \mathbb{N}} A_n$ .  $\square$

## Loeb Construction

In this section we provide the Loeb version of constructing a non-standard version of stochastic analysis (cf. [8]).

### Measure

Let a basic non-empty set  $\Omega$  be fixed. We recall that a sub-set  $\mathcal{A} \subset \mathbb{P}(\Omega)$  is called a Boolean algebra of sets, iff the following requirements are satisfied:

$$\begin{aligned} & \Omega \in \mathcal{A}; \\ & (\forall x \in \mathcal{A})(\forall y \in \mathcal{A}), \quad (x \cup y \in \mathcal{A}); \\ & (\forall x \in \mathcal{A})(\forall y \in \mathcal{A}), \quad (x \cap y \in \mathcal{A}); \\ & (\forall x \in \mathcal{A})(\forall y \in \mathcal{A}), \quad (x \setminus y \in \mathcal{A}). \end{aligned}$$

Or the last three sentences can be combined into one:

$$\begin{aligned} & (\forall x \in \mathcal{A})(\forall y \in \mathcal{A}), \\ & ((x \cup y \in \mathcal{A}) \wedge (x \cap y \in \mathcal{A}) \wedge (x \setminus y \in \mathcal{A})). \end{aligned}$$

If  $\mathcal{A}$  is closed under countable operations, it is then called a  $\sigma$ -algebra on  $\Omega$ :

$$\begin{aligned} & \forall (x_n: n \in \mathbb{N}), \\ & ((\forall n \in \mathbb{N}, x_n \in \mathcal{A}) \Rightarrow (\cup_{n \in \mathbb{N}} x_n \in \mathcal{A})). \end{aligned}$$

It is well-known that for any set  $\mathcal{B}$  of subsets of  $\Omega$ , there is uniquely a smallest  $\sigma$ -algebra  $\mathcal{A} \supset \mathcal{B}$ . This is the  $\sigma$ -algebra generated by  $\mathcal{B} \in \mathbb{P}(\mathbb{P}(\Omega))$ .

**Definition 44 (Probability Content and Measure)** Let  $\mathcal{A}$  be a Boolean algebra of sets on a total set  $\Omega$ . A function

$$\mu: A \in \mathcal{A} \mapsto \mu(A) \in \mathbb{R}$$

is called a content on  $(\Omega, \mathcal{A})$ , if the following conditions are satisfied:

$$\begin{aligned} (1) : & \forall x \in \mathcal{A}, \quad \mu(x) \geq 0; \\ (2) : & \forall x \in \mathcal{A}, \quad \forall y \in \mathcal{A}, (x \cap y = \emptyset) \\ & \Rightarrow (\mu(x \cup y) = \mu(x) + \mu(y)). \end{aligned}$$

If the following condition of “the continuity from above at the empty set” is satisfied, then  $\mu$  is called a measure:

$$\begin{aligned} & \forall (x_n: n \in \mathbb{N}), \quad ((\forall n \in \mathbb{N}, (x_n \in \mathcal{A})) \\ & \wedge (x_{n+1} \subset x_n) \wedge (\cap_{n \in \mathbb{N}} x_n = \emptyset)) \Rightarrow \left( \lim_{n \rightarrow \infty} \mu(x_n) = 0 \right). \end{aligned}$$

An equivalent condition is:  $\mu$  is  $\sigma$ -additive, in the sense that: if  $\langle A_n \rangle$  is a disjoint sequence in  $\mathcal{A}$ , with union  $\cup_{n \in \mathbb{N}} A_n \in \mathcal{A}$ , then

$$\mu(\cup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} \mu(A_n).$$

If content (or measure)  $\mu$  is normalized:  $\mu(\Omega) = 1$ , then it is called a probability content (or measure, respectively).

**Theorem 27 (Extension of a Measure)** If  $\mu$  is a measure on  $(\Omega, \mathcal{A})$ , and  $\mathcal{B}$  is the  $\sigma$ -algebra generated by  $\mathcal{A}$ , then there is a unique extension of  $\mu$  to a measure  $\bar{\mu}$  on  $(\Omega, \mathcal{B})$ . Further, let  $\bar{\mathcal{A}}$  be the collection of sets of the form  $A = (B \setminus D) \cup (D \setminus B)$ , where  $B \in \mathcal{B}$ , and  $D \subset C$ , with  $C \in \mathcal{B}$ ,  $\bar{\mu}(C) = 0$ , and define then  $\bar{\mu}(A) = \bar{\mu}(B)$ , then the definition is valid, and  $\bar{\mu}$  is a measure on  $(\Omega, \bar{\mathcal{A}})$ . This extension of  $\mu$  is the “completion of  $\mu$ ”. It is “complete” in the sense that whenever  $A \in \bar{\mathcal{A}}$ ,  $F \subset A$ , and  $\bar{\mu}(A) = 0$ , then  $F \in \bar{\mathcal{A}}$  with  $\bar{\mu}(F) = 0$ .

**Remark 8 ( $\sigma$ -Finiteness)** The definition of a content or measure, is generally extended to allow  $+\infty$  in the range. Then in the discussion the condition of  $\sigma$ -finiteness is generally sufficient and necessary: there is a sequence  $A_n \in \mathcal{A}$ , with  $\cup_{n \in \mathbb{N}} A_n = \Omega$ , and  $\mu(A_n)$  finite, for all  $n$ .

### Loeb Measure

**Definition 45 (Loeb Content)** Let  $\Omega \in {}^*S_\infty$  be an internal set,  $\mathcal{A}$  be an internal set of subsets of  $\Omega$ , which is closed under complementation and finite union. A Loeb content on  $(\Omega, \mathcal{A})$  is an internal function  $\mu: \mathcal{A} \Rightarrow {}^*\mathbb{R}$ , which is

a content on  $\mathcal{A}$ , i.e., it is finitely-additive and non-negatively valued. (But it is valued in  ${}^*\mathbb{R}$ , not real-valued.)

**Theorem 28 (Loeb Measure)** For a Loeb content space  $\langle \Omega, \mathcal{A}, \mu \rangle$ , define the standard-part of  $\mu$  as

$${}^\circ\mu: A \in \mathcal{A} \mapsto {}^\circ\mu(A) := {}^\circ(\mu(A)).$$

It is then  $\sigma$ -additive, and its completion is the Loeb measure  $L(\mu)$  of  $\mu$ , defined on the Loeb algebra  $L(\mathcal{A})$  consisting of those  $B \subset \Omega$  which are “ $\mu$ -approximable”: For any positive real  $\epsilon$ , there are  $A \in \mathcal{A}$ ,  $C \in \mathcal{A}$ , such that:

$$A \subset B \subset C, \quad \mu(C) - \mu(A) < \epsilon.$$

Indeed there is  $D \in \mathcal{A}$ , (called Loeb’s modification of  $B$ ), such that

$$L(\mu)(B \Delta D) = 0.$$

Here  $\Delta$  is the symmetric-difference:

$$B \Delta D := (B \setminus D) \cup (D \setminus B).$$

We omit the proof, but by the saturation property of internal sets, the continuity or  $\sigma$ -additivity of  $\mu$  is trivial. It should be emphasized that  $L(\mu)$  is a real-valued measure; therefore the integrands of its integration are real-valued, but defined on  $\Omega$ , hence is not internal.

**Definition 46 (Measurable Functions)** If temporarily we denote by  $\mathcal{G}$  the set of all open sets of  $\mathbb{R}$ , (also called the topology of  $\mathbb{R}$ ), and call elements of  ${}^*\mathcal{G}$  star-open, then an internal function  $F: \Omega \Rightarrow {}^*\mathbb{R}$  is called  $\mathcal{A}$ -measurable, iff:

$$(G \in {}^*\mathcal{G}) \Rightarrow (F^{-1}(G) \in \mathcal{A}).$$

If  $F := \pi_{\mathcal{U}}(F_j)$ ,  $\mu = \pi_{\mathcal{U}}(\mu_j)$ , we define naturally:

$${}^*\int F d\mu := \pi_{\mathcal{U}}\left(\int F_j d\mu_j\right).$$

This is valued in  ${}^*\mathbb{R}$ .

**Remark 9 (Loeb Integral)** If  $F$  is real-bounded, then

$${}^\circ\left({}^*\int F d\mu\right) = \int {}^\circ F dL(\mu),$$

which is then called the Loeb integral. The equality here is guaranteed by the  $S$ -integrability of the internal function  $F$ :

- (1):  ${}^*\int |F| d\mu$  is limited.
- (2): If  $\mu(A) = 0$ , then  ${}^*\int |F| d\mu = 0$ .
- (3): If  $F(\omega) = 0$  for all  $\omega \in A$ , then  ${}^*\int |F| d\mu \approx 0$ .

Now we turn to the integration of a real-valued function  $f$  on  $\Omega$ , with respect to the real-valued measure  $L(\mu)$ .

**Theorem 29 (Lifting of Function)** Any  $L(\mathcal{A})$ -measurable function  $f$  has a “lifting”  $F: \Omega \mapsto {}^*\mathbb{R}$ , in the sense that:  $F$  is internal,  $\mathcal{A}$ -measurable function, satisfying

$$L(\mu)\{\omega: {}^\circ F(\omega) \neq f(\omega)\} = 0.$$

**Definition 47 (Normed Counting Measure)** If  $\Omega \in {}^*\mathcal{S}_\infty$  is a hyperfinite set, and  $\mathcal{A}$  the set of all internal set, the internal probability content  $\mu$ ,

$$\mu(A) := \frac{{}^*\text{card}(A)}{{}^*\text{card}(\Omega)},$$

is called the normed counting measure on  $\Omega$ .

**Example 11 (Hyperfinite Equi-partition of the Unit-Interval)** Let  $\mathcal{N} \in {}^*\mathbb{N} \setminus \mathbb{N}$  be an unlimited hyperinteger,  $\mathcal{T}$  be the equi-partition hyperset of the unit interval into  $\mathcal{N}$  parts, and  $\mu$  be the normalized uniform counting measure. Then we get the Loeb measure  $L(\mu)$  and the Loeb algebra  $L(\mathcal{A})$  on  $\mathcal{T} \subset {}^*[0 \dots 1]$ . The Loeb measure space  $(\mathcal{T}, L(\mathcal{A}), L(\mu))$  is simply the pull-back of the standard Lebesgue measure space  $(\mathbb{U}, \mathcal{B}(\mathbb{U}), \lambda)$  by the standard-part map.

Here we have a dual of push-down and pull-back (or lifting). The push-down is via standard-part map  $\text{st}: \mathcal{T} \mapsto \mathbb{U}$ , which is a surjection, while its inverse  $\text{st}^{-1}$  furnishes the pull-back. The push-down of  $L(\mathcal{A})$ , by  $\text{st}$ , is by definition,

$$\{B \subset \mathbb{U}: \text{st}^{-1}[B] \in L(\mathcal{A})\}.$$

This is precisely the Lebesgue algebra  $\overline{\mathcal{B}(\mathbb{U})}$ . And for a member  $B$  in this latter algebra, then the push-down of  $L(\mu)$  will assign the measure value  $L(\mu)(\{x \in \mathcal{T}: {}^*x \in B\})$ . But we have already shown that when  $B$  is an interval, this value is simply its length  $\lambda(B)$ . This is enough to identify the pushed-down measure. Actually, the same kind of Loeb construction may be applied to the whole real axis  $\mathbb{R}$ , or the real half-axis  $\mathbb{R}_+$ , instead of  $\mathbb{U}$  (or the analogous finite intervals). A little bit of attention should be paid to the  $\sigma$ -finiteness, of course.

### Anderson’s Wiener Process

**Definition 48 (Hyperfinite Time-axis)** When  $\mathcal{N} \in {}^*\mathbb{N} \setminus \mathbb{N}$  is fixed,  $\mathcal{N} = \pi_{\mathcal{U}}(N_j)$ , we have the hyperfinite time-axis

$$\mathcal{T} := \pi_{\mathcal{U}}(T_j); T_j := \left\{ \frac{k}{N_j}: k \in \mathbb{Z}, 0 \leq k < N_j^2 \right\}.$$



As before,  $T_j$  is the interval  $[0 \dots N_j]$  equi-partitioned into  $N_j^2$  parts, each with length  $1/N_j$ . The right-end  $N_j$  is deliberately deleted, so that  $\text{card}(T_j) = N_j^2$ , while the augmented set is  $T'_j := T_j \cup \{N_j\}$ . The set  $T_j$  will be called rank  $k$  (discretized) time-axis. The left- and the right- approximations from the semi-axis  $[0 \dots +\infty)$  to  $\mathcal{T}$  is well-defined:

$$\begin{cases} \text{floor}_{\mathcal{T}}(c) = \pi_{\mathcal{U}} \left\langle \frac{\text{floor}(c * N_j)}{N_j} : j \in I \right\rangle; \\ \text{ceil}_{\mathcal{T}}(c) = \pi_{\mathcal{U}} \left\langle \frac{\text{ceil}(c * N_j)}{N_j} : j \in I \right\rangle. \end{cases}$$

**Definition 49 (Hyperfinite Coin-Tossing)** With  $\mathcal{N} \in {}^*\mathbb{N} \setminus \mathbb{N}$ , and  $\mathcal{T} = \left\{ \frac{k}{\mathcal{N}} : k = 0, 1, 2, \dots, \mathcal{N} - 1 \right\}$  as above, we define:

$$\begin{aligned} \text{Coin} &:= \{-1, +1\}; \\ \Omega_j &:= \text{Coin}^{T_j}; \\ \Omega &:= \pi_{\mathcal{U}} \langle \Omega_j \rangle; \end{aligned}$$

Of course,  $+1$  or  $-1$  means “head” or “tail” of a coin toss. Any element  $\omega_j \in \Omega_j$  is called a coin-particle of rank  $j$ , and an element  $\omega \in \Omega$  is called a coin-particle (of hyper-finite rank). The normalized counting measure  $P_j$  on  $\Omega_j$  is simple indeed, as is the one occurring most often in our high-school textbooks. (We will denote by  $\mathbb{E}_{P_j}$  the expectation with respect to  $P_j$ ). The normalized counting measure  $P$  is defined on the algebra of all internal subsets of the hyper-finite internal set  $\Omega$ . Therefore the Loeb construction give us the Loeb measure  $L(P)$ , called the hyperfinite coin-measure, on the Loeb algebra  $L(\mathcal{A})$ , which is called the hyper-finite coin algebra.

**Definition 50 (The (Rank  $j$ ) Random Walker)** For fixed  $j$ , and any rank  $j$  coin-particle  $\omega_j \in \text{Coin}^{T_j}$ , its path is the mapping  $t \mapsto B_j(\omega_j, t) \in \mathbb{R}$ , defined as follows. The domain of definition is

$$T'_j := \left\{ \frac{k}{N_j} : k = 0, 1, 2, \dots, N_j^2 \right\}.$$

And when  $t = \frac{k}{N_j}$ ,

$$B_j(\omega_j, t) := \frac{1}{\sqrt{N_j}} \sum \left\{ \omega_j \left( \frac{m}{N_j} \right) : 0 \leq m < k \right\}.$$

By definition, the walk starts from the origin  $B_j(\omega_j, 0) := 0$ . Then recursively, at each time  $t = m/N_j$ , when the ( $j$ -) particle is located at the position  $B_j(\omega_j, t)$ , the coin-tossing  $\omega_j(t)$  will decide the plus or minus direction of the next

step, occurring in the time-interval, from  $t$  to  $t + \Delta_j t$ . Here the rank  $j$

$$\begin{aligned} \text{time-unit is } \Delta_j t &:= \frac{1}{N_j}; \\ \text{space-unit is } \Delta_j x &:= \frac{1}{\sqrt{N_j}}. \end{aligned}$$

As a coin-particle  $\omega_j \in \text{Coin}^{T_j}$  is randomly chosen, we may call it a random walker.

**Definition 51 (Hyperfinite Random Walk)** Define entities in the hyper-universe:

$$\begin{aligned} B &:= \pi_{\mathcal{U}} \langle B_j \rangle; \\ B(\omega, t) &:= \sum \left\{ \frac{\omega(s)}{\sqrt{\mathcal{N}}} : s < t \right\}; \end{aligned}$$

$B: \langle \omega, t \rangle \in \Omega \times \mathcal{T} \mapsto B(\omega, t) \in {}^*\mathbb{R}$  is called the hyper-finite random walk.

**Lemma 8 (Independent Increment (Finite-Rank))** For fixed  $j$ , if there are  $2n$  elements of  $T'_j$ , arranged according to order:

$$s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_n < t_n.$$

Then these  $n$  increments:

$$\begin{aligned} B_j(\cdot, t_1) - B_j(\cdot, s_1), B_j(\cdot, t_2) - B_j(\cdot, s_2), \\ \dots, B_j(\cdot, t_n) - B_j(\cdot, s_n), \end{aligned}$$

are independent, in the sense that: for any choice of intervals (or Borel sets)  $A_1, A_2, \dots, A_n$ ,

$$\begin{aligned} P_j \{ \omega_j : B_j(\omega_j, t_k) - B_j(\omega_j, s_k) \in A_k, \forall k = 1, 2, \dots, n, \} \\ = \prod_{1 \leq k \leq n} P_j \{ \omega_j : B_j(\omega_j, t_k) - B_j(\omega_j, s_k) \in A_k \}. \end{aligned}$$

*Proof* Coin-tossings are independent, by assumption. By transference,  $\square$

**Lemma 9 (Independent Increment (Hyperfinite-Rank))** If  $2n$  elements of  $\mathcal{T}$  are chosen and arranged according to order:

$$s_1 < t_1 \leq s_2 < t_2 \leq \dots \leq s_n < t_n.$$

Then the increments:

$$\begin{aligned} B(\cdot, t_1) - B(\cdot, s_1), B(\cdot, t_2) - B(\cdot, s_2), \\ \dots, B(\cdot, t_n) - B(\cdot, s_n), \end{aligned}$$

are  $\star$ -independent, in the sense that: for any choice of internal sets  $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$ ,

$$\begin{aligned} P\{\omega: B(\omega, t_k) - B(\omega, s_k) \in \tilde{A}_k, \text{ when } k = 1, 2, \dots, n\} \\ = \prod_{1 \leq k \leq n} P\{\omega: B(\omega, t_k) - B(\omega, s_k) \in \tilde{A}_k\}. \end{aligned}$$

**Lemma 10 (Fourier and Moments Characteristic, Finite-Rank)** For  $t \in T_j$ , and  $y \in \mathbb{R}$ ,

$$\begin{aligned} \mathbb{E}_{P_j} \left( e^{\sqrt{-1}\theta * B_j(t)} \right) &= \left( \cos \left( \frac{y}{\sqrt{N_j}} \right) \right)^{\frac{t}{N_j}}; \\ \mathbb{E}_{P_j}(B_j(t)) &= 0; \\ \mathbb{E}_{P_j}(B_j(t))^2 &= t; \\ \mathbb{E}_{P_j}(B_j(t))^4 &= 3t^2 - 2\frac{t}{N_j}. \end{aligned}$$

*Proof* To calculate  $\mathbb{E}_{P_j} \left( e^{\sqrt{-1}\theta * B_j(t)} \right)$ , by independence, we need only calculate the case of only one-step, i. e., for  $t = \frac{1}{N_j}$ . Then it is

$$\frac{1}{2} \left( e^{\sqrt{-1}\frac{y}{\sqrt{N_j}}} + e^{\sqrt{-1}\frac{-y}{\sqrt{N_j}}} \right) = \cos \left( \frac{y}{\sqrt{N_j}} \right).$$

The odd moments are zero for the even-symmetry of the distribution. Then we do, for the 2nd moment,

$$(u_1 + u_2 + u_3 + \dots)^2 = \sum u_k^2 + 2 \sum_{k < l} u_k u_l.$$

Also: for the 4th moment,

$$(u_1 + u_2 + \dots)^4 = \sum u_k^4 + 6 \sum_{k < l} u_k^2 u_l^2 + \text{odd terms}.$$

By transference, then:  $\square$

**Lemma 11 (Fourier and Moments Characteristic, Hyperfinite-Rank)** For any  $t \in \mathcal{T}$ , and for any  $y \in \mathbb{R}$ ,

$$\begin{aligned} \mathbb{E}_P \left( e^{\sqrt{-1}y * B(t)} \right) &= \left( \cos \left( \frac{y}{\sqrt{\mathcal{N}}} \right) \right)^{\frac{t}{\mathcal{N}}}; \\ \mathbb{E}_P(B(t)) &= 0; \\ \mathbb{E}_P(B(t))^2 &= t; \\ \mathbb{E}_P(B(t))^4 &= 3t^2 - 2\frac{t}{\mathcal{N}}; \end{aligned}$$

**Theorem 30 (Anderson)** The function  $b$  defined below is a Wiener process.

$$b(\omega, t) := \text{st}(B(\omega, \text{ceil}_{\mathcal{T}}(t))); \quad b: \Omega \times [0 \dots \infty) \mapsto \mathbb{R}.$$

• As to the independence of the increments:

$$\begin{aligned} b(\cdot, t_1) - b(\cdot, s_1), \quad b(\cdot, t_2) - b(\cdot, s_2), \dots, \\ b(\cdot, t_n) - b(\cdot, s_n), \end{aligned}$$

this utilizes the Loeb's modification, and the standard-part transformation between  $b$  and  $B$ .

• As to the Gaussian nature of the increment  $b(\cdot, t) - b(\cdot, s)$ , which is distributed just like  $b(\cdot, t - s)$ , calculation of its Fourier characteristic is the most convenient way of proof. Now we have (by calculus, L'Hospital or not):

$$\lim_{n \rightarrow \infty} \cos^{\text{floor}(nt)} \left( \frac{y}{\sqrt{n}} \right) = e^{-\frac{y^2}{2}t}.$$

Therefore

$$\int e^{\sqrt{-1}yb(t)} dL(P) = e^{-\frac{y^2}{2}t}.$$

• As to the continuity of sample-paths, the usual method is to calculate the moments of increment. The discussion we will omit here.

## A Future for Non-standard Analysis?

The ultrapower construction above is the semantic approach. Another approach is called the *syntactic approach* to non-standard analysis requires much less logic and model theory to understand and use. This approach was developed in the mid-1970s by the mathematician Edward Nelson [9,10], using a theory called the *internal set theory* which has a unary operation “standard” that makes non-standard analysis possible. Despite the elegance exhibited in non-standard analysis, some claim that it is not very useful and can only do reinterpretations or reproofs of previous results like [4,14]. In the study of random walks and such, NSA still find some applications, such as [1].

## Bibliography

1. Albeverio S, Fenstad JE, Hoegh-Krohn R, Lindström T (1986) Nonstandard Methods in Stochastic Analysis and Mathematical Physics. Bull Am Math Soc 17(2):385–389
2. Bernstein A, Robinson A (eds) (1966) Solution of an invariant subspace problem of KT Smith and PR Halmos. Pacific J Math 16(3):421–431
3. Halmos P (1966) Invariant subspaces for Polynomially Compact Operators. Pacific J Math 16(3):433–437
4. Kamae T (1982) A simple proof of the ergodic theorem using nonstandard analysis. Israel J Math 42(4):284–290
5. Keisler HJ (1986) An Infinitesimal Approach to Stochastic Analysis. J Symb Logic 51(3):822–824 now included In: (1984) Mem Amer Math Soc 297

6. Keisler HJ (1986) Elementary Calculus: An Approach Using Infinitesimals, 2nd edn. Now available at <http://www.math.wisc.edu/~keisler/calc.html>
7. Kelley JL (1975) General Topology. Springer, New York
8. Loeb PA, Wolff MPH (2000) Nonstandard Analysis for the Working Mathematician. Math. Appl., vol. 510; Kluwer, Dordrecht
9. Nelson E (1977) Internal Set Theory A New Approach to Non-standard Analysis. Bull Am Math Soc 83(6):1165–1198; Also see <http://www.math.princeton.edu/~nelson/books/1.pdf>
10. Nelson E (1989) Radically Elementary Probability Theory. Bull Am Math Soc 20(2):240–243
11. Robert A (1985) Nonstandard Analysis. Bull Am Math Soc 16(2):298–306
12. Robinson A (1966) Non-standard analysis, rev edn. North Holland, Amsterdam
13. Schmieden C, Laugwitz D (1958) Eine Erweiterung der Infinitesimalrechnung. Math Zeit 69:1–39
14. van den Dries L, Wilkie AJ (1984) Gromov's Theorem on Groups of Polynomial Growth and Elementary Logic. J Algebra 89:349–374

## Normal Forms in Perturbation Theory

HENK W. BROER

University of Groningen, Groningen, The Netherlands

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Motivation  
 The Normal Form Procedure  
 Preservation of Structure  
 Semi-local Normalization  
 Non-formal Aspects  
 Applications  
 Future Directions  
 Bibliography

### Glossary

**Normal form procedure** This is the stepwise ‘simplification’ by changes of coordinates, of the Taylor series at an equilibrium point, or of similar series at periodic or quasi-periodic solutions.

**Preservation of structure** The normal form procedure is set up in such a way that all coordinate changes preserve a certain appropriate structure. This applies to the class of **Hamiltonian** or **volume preserving** systems, as well as to systems that are **equivariant** or **reversible** with respect to a symmetry group. In all cases the systems may also depend on **parameters**.

**Symmetry reduction** The truncated normal form often exhibits a toroidal symmetry that can be factored out, thereby leading to a lower dimensional reduction.

**Perturbation theory** The attempt to extend properties of the (possibly reduced) normal form truncation, to the full system.

### Definition of the Subject

Nonlinear dynamical systems are notoriously hard to tackle by analytic means. One of the few approaches that has been effective for the last couple of centuries, is Perturbation Theory. Here systems are studied, which in an appropriate sense, can be seen as perturbations of a given system with ‘well-known’ dynamical properties. Such ‘well-known’ systems usually are systems with a great amount of symmetry (like integrable Hamiltonian systems [1]) or very low-dimensional systems. The methods of Perturbation Theory then try to extend the ‘well-known’ dynamical properties to the perturbed system. Methods to do this are often based on the Implicit Function Theorem, on normal hyperbolicity [85,87] or on Kolmogorov–Arnold–Moser theory [1,10,15,18,38].

To obtain a perturbation theory set-up, normal form theory is a vital tool. In its most elementary form it amounts to ‘simplifying’ the Taylor series of a dynamical system at an equilibrium point by successive changes of coordinates. The lower order truncation of the series often belongs to the class of ‘well-known’ systems and by the Taylor formula the original system then locally can be viewed as a perturbation of this ‘well-known’ truncation, that thus serves as the ‘unperturbed’ system. More involved versions of normal form theory exist at periodic or quasi-periodic evolutions of the dynamical system.

### Introduction

We review the formal theory of normal forms of dynamical systems at equilibrium points. Systems with continuous time, i.e., of vector fields, or autonomous systems of ordinary differential equations, are considered extensively. The approach is universal in the sense that it applies to many cases where a structure is preserved. In that case also the normalizing transformations preserve this structure. In particular this includes the Hamiltonian and the volume preserving case as well as cases that are equivariant or reversible with respect to a symmetry group. In all situations the systems may depend on parameters. Related topics are being dealt with concerning a vector field at a periodic solution or a quasi-periodic invariant torus as well as the case of a diffeomorphism at a fixed point. The paper

is concluded by discussing a few non-formal aspects and some applications.

### Motivation

The term ‘normal form’ is widely used in mathematics and its meaning is very sensitive for the context. In the case of linear maps from a given vector space to itself, for example, one may consider all possible choices of a basis. Each choice gives a matrix-representation of a given linear map. A suitable choice of basis now gives the well-known Jordan canonical form. This normal form, in a simple way displays certain important properties of the linear map, concerning its spectrum, its eigenspaces, and so on.

Presently the concern is with dynamical systems, such as vector fields (i. e., systems of autonomous ordinary differential equations), or diffeomorphisms. The aim is to simplify these systems near certain equilibria and (quasi-periodic) solutions by a proper choice of coordinates. The purpose of this is to better display the local dynamical properties. This normalization is effected by a stepwise simplification of formal power series (such as Taylor series).

### Reduction of Toroidal Symmetry

In a large class of cases, the simplification of the Taylor series induces a toroidal symmetry up to a certain order, and truncation of the series at that order gives a local approximation of the system at hand. From this truncation we can reduce this toroidal symmetry, thereby also reducing the dimension of the phase space. This kind of procedure is reminiscent of the classical reductions in classical mechanics related to the Noether Theorem, compare with [1],

#### ► Dynamics of Hamiltonian Systems.

This leads us to a first perturbation problem to be considered. Indeed, by truncating and factoring out the torus symmetry we get a polynomial system on a reduced phase, and one problem is how persistent this system is with respect to the addition of higher order terms. In the case where the system depends on parameters, the persistence of a corresponding bifurcation set is of interest.

In many examples the reduced phase space is 2-dimensional where the dynamics is qualitatively determined by a polynomial and this perturbation problem can be handled by Singularity Theory. In quite a number of cases the truncated and reduced system turns out to be structurally stable [67,86].

### A Global Perturbation Theory

When returning to the original phase space we consider the original system as a perturbation of the de-reduced

truncation obtained so far. In the ensuing perturbation problem, several types of resonance can play a role.

A classical example [2,38] occurs when in the reduced model a Hopf bifurcation takes place and we have reduced by a 1-torus. Then in the original space generically a Hopf–Neimark–Sacker bifurcation occurs, where in the parameter space the dynamics is organized by resonance *tongues*. In cases where we have reduced by a torus of dimension larger than 1, we are dealing with quasi-periodic Hopf bifurcation, in which the bifurcation set gets ‘Cantorized’ by Diophantine conditions, compare with [6,15,21,38]. Also the theory of homo- and heteroclinic bifurcations then is of importance [68].

We use the two perturbation problems as a motivation for the Normal Form Theory to be reviewed, for details, however, we just refer to the literature. For example, compare with ► [Perturbation Theory](#) and references therein.

### The Normal Form Procedure

The subject of our interest is the simplification of the Taylor series of a vector field at a certain equilibrium point. Before we explore this, however, let us first give a convenient normal form of a vector field near a non-equilibrium point.

**Theorem 1 (Flow box [81])** *Let the  $C^\infty$  vector field  $X$  on  $\mathbb{R}^n$  be given by  $\dot{x} = f(x)$  and assume that  $f(p) \neq 0$ . Then there exists a neighborhood of  $p$  with local  $C^\infty$  coordinates  $y = (y_1, y_2, \dots, y_n)$ , such that in these coordinates  $X$  has the form*

$$\begin{aligned}\dot{y}_1 &= 1 \\ \dot{y}_j &= 0,\end{aligned}$$

for  $2 \leq j \leq n$ .

Such a local chart usually is called a *flowbox* and the above theorem the Flowbox Theorem. A proof simply can be given using a transversal local  $C^\infty$  section that cuts the flow of the vector field transversally. For the coordinate  $y_1$  then use the time-parametrization of the flow, while the coordinates  $y_j$ , for  $2 \leq j \leq n$  come from the section, compare with, e. g., [81].

### Background, Linearization

The idea of simplification near an equilibrium goes back at least to Poincaré [69], also compare with Arnold [2]. To be definite, we now let  $X$  be a  $C^\infty$  vector field on  $\mathbb{R}^n$ , with the origin as an equilibrium point. Suppose that  $X$  has the form  $\dot{x} = Ax + f(x)$ ,  $x \in \mathbb{R}^n$ , where  $A$  is linear and

where  $f(0) = 0$ ,  $D_0 f = 0$ . The first idea is to apply successive  $C^\infty$  changes of coordinates of the form  $\text{Id} + P$ , with  $P$  a homogeneous polynomial of degree  $m = 2, 3, \dots$ , ‘simplifying’ the Taylor series step by step.

The most ‘simple’ form that can be obtained in this way, is where *all* higher order terms vanish. In that case the normal form is formally linear. Such a case was treated by Poincaré, and we shall investigate this now.

We may even assume to work on  $\mathbb{C}^n$ , for simplicity assuming that the eigenvalues of  $A$  are distinct. A collection  $\lambda = (\lambda_1, \dots, \lambda_n)$  of points in  $\mathbb{C}$  is said to be *resonant* if there exists a relation of the form

$$\lambda_s = \langle r, \lambda \rangle,$$

for  $r = (r_1, \dots, r_n) \in \mathbb{Z}^n$ , with  $r_k \geq 0$  for all  $k$  and with  $\sum r_k \geq 2$ . The *order* of the resonance then is the number  $|r| = \sum r_k$ . The Poincaré Theorem now reads

**Theorem 2 (Formal linearization [2,69])** *If the (distinct) eigenvalues  $\lambda_1, \dots, \lambda_n$  of  $A$  have no resonances, there exists a formal change of variables  $x = y + O(|y|^2)$ , transforming the above vector field  $X$ , given by*

$$\dot{x} = Ax + f(x)$$

to

$$\dot{y} = Ay.$$

We include a proof [2], since this will provide the basis for almost all further considerations.

*Proof* The formal power series  $x = y + O(|y|^2)$  is obtained in an inductive manner. Indeed, for  $m = 2, 3, \dots$  a polynomial transformation  $x = y + P(y)$  is constructed, with  $P$  homogeneous of degree  $m$ , which removes the terms of degree  $m$  from the vector field. At the end we have to take the composition of all these polynomial transformations.

1. The basic tool for the  $m$ th step is the following. Let  $v$  be homogeneous of degree  $m$ , then if the vector fields  $\dot{x} = Ax + v(x) + O(|x|^{m+1})$  and  $\dot{y} = Ay$  are related by the transformation  $x = y + P(y)$  with  $P$  also homogeneous of degree  $m$ , then

$$D_x P A x - A P(x) = v(x).$$

This relation usually is called the *homological equation*, the idea being to determine  $P$  in terms of  $v$ : by this choice of  $P$  the term  $v$  can be transformed away.

The proof of this relation is straightforward. In fact,

$$\begin{aligned} \dot{x} &= (\text{Id} + D_y P) A y \\ &= (\text{Id} + D_y P) A (x - P(x) + O(|x|^{m+1})) \\ &= A x + \{D_x P A x - A P(x)\} + O(|x|^{m+1}), \end{aligned}$$

where we used that for the inverse transformation we know  $y = x - P(x) + O(|x|^{m+1})$ .

2. For notational convenience we introduce the linear operator  $\text{ad}A$ , the so-called adjoint operator, by

$$\text{ad}A(P)(x) := D_x P A x - A P(x),$$

then the homological equation reads  $\text{ad}A(P) = v$ . So the question is reduced to whether  $v$  is in the image of the operator  $\text{ad}A$ . It turns out that the eigenvalues of  $\text{ad}A$  can be expressed in those of  $A$ . If  $x_1, x_2, \dots, x_n$  are the coordinates corresponding to the basis  $e_1, e_2, \dots, e_n$ , again it is a straightforward computation to show that for  $P(x) = x^r e_s$ , one has

$$\text{ad}A(P) = (\langle r, \lambda \rangle - \lambda_s) P.$$

Here we use the multi-index notation  $x^r = x_1^{r_1} x_2^{r_2} \dots x_n^{r_n}$ . Indeed, for this choice of  $P$  one has  $A P(x) = \lambda_s P(x)$ , while

$$\frac{\partial x^r}{\partial x} A x = \sum_j \frac{r_j}{x_j} x^r \lambda_j x_j = \langle r, \lambda \rangle x^r.$$

We conclude that the monomials  $x^r e_s$  are eigenvectors corresponding to the eigenvalues  $\langle r, \lambda \rangle - \lambda_s$ . We conclude that the operator  $\text{ad}A$  is *semisimple*, since it has a basis of eigenvectors. Therefore, if  $\ker \text{ad}A = 0$ , the operator is surjective. This is exactly what the non-resonance condition on the eigenvalues of  $A$  amounts to. More precisely, the homological equation  $\text{ad}A(P) = v$  can be solved for  $P$  for each homogeneous part  $v$  of degree  $m$ , provided that there are no resonances up to order  $m$ .

3. The induction process now runs as follows. For  $m \geq 2$ , given a form

$$\dot{x} = A x + v_m(x) + O(|x|^{m+1}),$$

we solve the homological equation

$$\text{ad}A(P_m) = v_m,$$

then carrying out the transformation  $x = y + P_m(y)$ . This takes the above form to

$$\dot{y} = A y + v_{m+1}(y) + O(|y|^{m+2}).$$

The composition of all the polynomial transformations then gives the desired formal transformation.  $\square$



*Remark*

- It is well-known that the formal series usually diverge. Here we do not go into this problem, for a brief discussion see below.
- If resonances are excluded up to a finite order  $N$ , we can linearize up to that order, so obtaining a normal form.

$$\dot{y} = Ay + O(|y|^{N+1}).$$

In this case the transformation can be taken as a polynomial.

- If the original problem is real, but with the matrix  $A$  having non-real eigenvalues, we still can keep all transformations real by also considering complex conjugate eigenvectors.

We conclude this introduction by discussing two further linearization theorems, one due to Sternberg and the other to Hartman–Grobman. We recall the following for a vector field  $X(x) = Ax + f(x)$ ,  $x \in \mathbb{R}^n$ , with 0 as an equilibrium point, i. e., with  $f(0) = 0$ ,  $D_0 f = 0$ . The equilibrium 0 is *hyperbolic* if the matrix  $A$  has no purely imaginary eigenvalues. Sternberg's Theorem reads

**Theorem 3 (Smooth linearization [82])** *Let  $X$  and  $Y$  be  $C^\infty$  vector fields on  $\mathbb{R}^n$ , with 0 as a hyperbolic equilibrium point. Also suppose that there exists a formal transformation  $(\mathbb{R}^n, 0) \rightarrow (\mathbb{R}^n, 0)$  taking the Taylor series of  $X$  at 0 to that of  $Y$ . Then there exists a local  $C^\infty$ -diffeomorphism  $\Phi : (\mathbb{R}^n, 0) \rightarrow (\mathbb{R}^n, 0)$ , such that  $\Phi_* X = Y$ .*

We recall that  $\Phi_* X(\Phi(x)) = D_x \Phi X(x)$ . This means that  $X$  and  $Y$  are locally *conjugated* by  $\Phi$ , the evolution curves of  $X$  are mapped to those of  $Y$  in a time-preserving manner. In particular Sternberg's Theorem applies when the conclusion of Poincaré's Theorem holds: for  $Y$  just take the linear part  $Y(x) = Ax(\partial/\partial x)$ .

Combining these two theorems we find that in the hyperbolic case, under the exclusion of all resonances, the vector field  $X$  is linearizable by a  $C^\infty$ -transformation. The Hartman–Grobman Theorem moreover says that the non-resonance condition can be omitted, provided we only want a  $C^0$ -linearization.

**Theorem 4 (Continuous linearization e. g., [67])** *Let  $X$  be a  $C^\infty$  vector field on  $\mathbb{R}^n$ , with 0 as a hyperbolic equilibrium point. Then there exists a local homeomorphism  $\Phi : (\mathbb{R}^n, 0) \rightarrow (\mathbb{R}^n, 0)$ , locally conjugating  $X$  to its linear part.*

**Preliminaries from Differential Geometry**

Before we develop a more general Normal Form Theory we recall some elements from differential geometry. One

central notion used here is that of the *Lie derivative*. For simplicity all our objects will be of class  $C^\infty$ . Given a vector field  $X$  we can take any tensor  $\tau$  and define its Lie-derivative  $\mathcal{L}_X \tau$  with respect to  $X$  as the infinitesimal transformation of  $\tau$  along the flow of  $X$ . In this way  $\mathcal{L}_X \tau$  becomes a tensor of the same type as  $\tau$ . To be more precise, for  $\tau$  a real function  $f$  one so defines

$$\mathcal{L}_X f(x) = X(f)(x) = df(X)(x) = \left. \frac{d}{dt} \right|_{t=0} f(X_t(x)),$$

i. e., the directional derivative of  $f$  with respect to  $X$ . Here  $X_t$  denotes the flow of  $X$  over time  $t$ . For  $\tau$  a vector field  $Y$  one similarly defines

$$\begin{aligned} \mathcal{L}_X Y(x) &= \left. \frac{d}{dt} \right|_{t=0} (X_{-t})_* Y(x) \\ &= \lim_{t \rightarrow 0} \frac{1}{t} \{ (X_{-t})_* Y(x) - Y(x) \} \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \{ Y(x) - (X_h)_* Y(x) \}, \end{aligned}$$

and similarly for *differential forms*, etc. Another central notion is the *Lie-brackets*  $[X, Y]$  defined for any two vector fields  $X$  and  $Y$  on  $\mathbb{R}^n$  by

$$[X, Y](f) = X(Y(f)) - Y(X(f)).$$

Here  $f$  is any real function on  $\mathbb{R}^n$  while, as before,  $X(f)$  denotes the directional derivative of  $f$  with respect to  $X$ .

We recall the expression of Lie-brackets in coordinates. If  $X$  is given by the system of differential equations  $\dot{x}_j = X_j(x)$ , with  $1 \leq j \leq n$ , then the directional derivative  $X(f)$  is given by  $X(f) = \sum_{j=1}^n X_j \partial f / \partial x_j$ . Then, if  $[X, Y]$  is given by the system  $\dot{x}_j = Z_j(x)$ , for  $1 \leq j \leq n$ , of differential equations, one directly shows that

$$Z_j = \sum_{k=1}^n \left( X_k \frac{\partial Y_j}{\partial x_k} - Y_k \frac{\partial X_j}{\partial x_k} \right).$$

Here  $Y_j$  relates to  $Y$  as  $X_j$  does to  $X$ . We list some useful properties.

**Proposition 5 (Properties of the Lie-derivative [81])**

1.  $\mathcal{L}_X(Y_1 + Y_2) = \mathcal{L}_X Y_1 + \mathcal{L}_X Y_2$  (linearity over  $\mathbb{R}$ )
2.  $\mathcal{L}_X(fY) = X(f) \times Y + f \times \mathcal{L}_X Y$  (Leibniz rule)
3.  $[Y, X] = -[X, Y]$  (skew symmetry)
4.  $[[X, Y], Z] + [[Z, X], Y] + [[Y, Z], X] = 0$  (Jacobi identity)
5.  $\mathcal{L}_X Y = [X, Y]$
6.  $[X, Y] = 0 \Leftrightarrow X_t \circ Y_s = Y_s \circ X_t$

*Proof* The first four items are left to the reader.

5. The equality  $\mathcal{L}_X Y = [X, Y]$  can be proven by observing the following. Both members of the equality are defined intrinsically, so it is enough to check it in any choice of (local) coordinates. Moreover, we can restrict our attention to the set  $\{x | X(x) \neq 0\}$ . By the Flowbox Theorem 1 we then may assume for the component functions of  $X$  that  $X_1(x) = 1$ ,  $X_j(x) = 0$  for  $2 \leq j \leq n$ . It is easy to see that both members of the equality now are equal to the vector field  $Z$  with components

$$Z_j = \frac{\partial Y_j}{\partial x_1}.$$

6. Remains the equivalence of the commuting relationships. The commuting of the flows, by a very general argument, implies that the bracket vanishes. In fact, fixing  $t$  we see that  $X_t$  conjugates the flow of  $Y$  to itself, which is equivalent to  $(X_t)_* Y = Y$ . By definition this implies that  $\mathcal{L}_X Y = 0$ .

Conversely, let  $c(t) = ((X_t)_* Y)(p)$ . From the fact that  $\mathcal{L}_X Y = 0$  it then follows that  $c(t) \equiv c(0)$ . Observe that the latter assertion is sufficient for our purposes, since it implies that  $(X_t)_* Y = Y$  and therefore  $X_t \circ Y_s = Y_s \circ X_t$ . Finally, that  $c(t)$  is constant can be shown as follows.

$$\begin{aligned} c'(t) &= \lim_{h \rightarrow 0} \frac{1}{h} \{c(t+h) - c(t)\} \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \{((X_{t+h})_* Y)(p) - ((X_t)_* Y)(p)\} \\ &= (X_t)_* \lim_{h \rightarrow 0} \frac{1}{h} \{((X_h)_* Y)(X_{-t}(p)) - Y(X_{-t}(p))\} \\ &= (X_t)_* \mathcal{L}_X Y(X_{-t}(p)) \\ &= (X_t)_*(0) \\ &= 0. \end{aligned}$$

□

### ‘Simple’ in Terms of an Adjoint Action

We now return to the setting of Normal Form Theory at equilibrium points. So given is the vector field  $\dot{x} = X(x)$ , with  $X(x) = Ax + f(x)$ ,  $x \in \mathbb{R}^n$ , where  $A$  is linear and where  $f(0) = 0$ ,  $D_0 f = 0$ . We recall that it is our general aim to ‘simplify’ the Taylor series of  $X$  at 0.

However, we have not yet said what the word ‘simple’ means in the present setting. In order to understand what is going on, we reintroduce the adjoint action associated to the linear part  $A$ , defined on the class of all  $C^\infty$  vector fields on  $\mathbb{R}^n$ . To be precise, this adjoint action  $\text{ad}A$  is defined by the Lie-bracket

$$\text{ad}A: Y \mapsto [A, Y],$$

where  $A$  is identified with the linear vector field  $\dot{x} = Ax$ . It is easily seen that this fits with the notation introduced in Theorem 2.

Let  $H^m(\mathbb{R}^n)$  denote the space of polynomial vector fields, homogeneous of degree  $m$ . Then the Taylor series of  $X$  can be viewed as an element of the product  $\prod_{m=1}^{\infty} H^m(\mathbb{R}^n)$ . Also, it directly follows that  $\text{ad}A$  induces a linear map  $H^m(\mathbb{R}^n) \rightarrow H^m(\mathbb{R}^n)$ , to be denoted by  $\text{ad}_m A$ . Let

$$B^m := \text{im } \text{ad}_m A,$$

the image of the map  $\text{ad}_m A$  in  $H^m(\mathbb{R}^n)$ . Then for *any* complement  $G^m$  of  $B^m$  in  $H^m(\mathbb{R}^n)$ , in the sense that

$$B^m \oplus G^m = H^m(\mathbb{R}^n),$$

we define the *corresponding* notion of ‘simplicity’ by requiring the homogeneous part of degree  $m$  to be in  $G^m$ . In the case of the Poincaré Theorem 2, since  $B^m = H^m(\mathbb{R}^n)$ , we have  $G^m = \{0\}$ .

We now quote a theorem from Sect. 7.6.1. in Du-mortier et al. [16]. Although its proof is very similar to the one of Theorem 2, we include it here, also because of its format.

**Theorem 6 (‘Simple’ in terms of  $G^m$  [74,82])** *Let  $X$  be a  $C^\infty$  vector field, defined in the neighborhood of  $0 \in \mathbb{R}^n$ , with  $X(0) = 0$  and  $D_0 X = A$ . Also let  $N \in \mathbb{N}$  be given and, for  $m \in \mathbb{N}$ , let  $B^m$  and  $G^m$  be such that  $B^m \oplus G^m = H^m(\mathbb{R}^n)$ . Then there exists, near  $0 \in \mathbb{R}^n$ , an analytic change of coordinates  $\Phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ , with  $\Phi(0) = 0$ , such that*

$$\Phi_* X(y) = Ay + g_2(y) + \cdots + g_N(y) + O(|y|^{N+1}),$$

with  $g_m \in G^m$ , for all  $m = 2, 3, \dots, N$ .

*Proof* We use induction on  $N$ . Let us assume that

$$X(x) = Ax + g_2(x) + \cdots + g_{N-1}(x) + f_N(x) + O(|x|^{N+1}),$$

with  $g_m \in G^m$ , for all  $m = 2, 3, \dots, N-1$  and with  $f_N$  homogeneous of degree  $N$ .

We consider a coordinate change  $x = y + P(y)$ , where  $P$  is polynomial of degree  $N$ , see above. For any such  $P$ , by substitution we get

$$\begin{aligned} (\text{Id} + D_y P)\dot{y} &= A(y + P(y)) + g_2(y) + \cdots \\ &\quad + g_{N-1}(y) + f_N(y) + O(|y|^{N+1}), \end{aligned}$$

or

$$\begin{aligned}\dot{y} &= (\text{Id} + D_y P)^{-1} (A(y + P(y)) + g_2(y) \\ &\quad + \cdots + g_{N-1}(y) + f_N(y) + O(|y|^{N+1})) \\ &= Ay + g_2(y) + \cdots + g_{N-1}(y) + f_N(y) + AP(y) \\ &\quad - D_y P A y + O(|y|^{N+1}),\end{aligned}$$

using that  $(\text{Id} + D_y P)^{-1} = \text{Id} - D_y P + O(|y|^N)$ . We conclude that the terms up to order  $N - 1$  are unchanged by this transformation, while the  $N$ th order term becomes

$$f_N(y) - \text{ad}_N A(P)(y).$$

Clearly, a suitable choice of  $P$  will put this term in  $G^N$ . This is the present version of the *homological equation* as introduced before.  $\square$

*Remark* For simplicity the formulations all are in the  $C^\infty$ -context, but obvious changes can be made for the case of finite differentiability. The latter case is of importance for applications of Normal Form Theory after reduction to a center manifold [85,87].

### Torus Symmetry

As a special case let the linear part  $A$  be semisimple, in the sense that it is diagonalizable over the complex numbers. It then directly follows that also  $\text{ad}_m A$  is semisimple, which implies that

$$\text{im } \text{ad}_m A \oplus \ker \text{ad}_m A = H^m(\mathbb{R}^n).$$

The reader is invited to provide the eigenvalues of  $\text{ad}_m A$  in terms of those of  $A$ , compare with the proof of Theorem 2. In the present case the obvious choice for the complementary space defining ‘simplicity’ is  $G^m = \ker \text{ad}_m A$ . Moreover, the fact that the normalized, viz. simplified, terms  $g_m$  are in  $G^m$  by definition means that

$$[A, g_m] = 0.$$

This, in turn, implies that  $N$ -jet of  $\Phi_\star X$ , i. e., the *normalized part* of  $\Phi_\star X$ , by Proposition 5 is invariant under all linear transformations

$$\exp tA, \quad t \in \mathbb{R}$$

generated by  $A$ . For further reading also compare with Sternberg [82], Takens [84], Broer [7,8] or [12].

More generally, let  $A = A_s + A_n$  be the Jordan canonical splitting in the semisimple and nilpotent part. Then one directly shows that  $\text{ad}_m A = \text{ad}_m A_s + \text{ad}_m A_n$  is the

Jordan canonical splitting, whence, by a general argument [89] it follows that

$$\text{im } \text{ad}_m A + \ker \text{ad}_m A_s = H^m(\mathbb{R}^n),$$

so where the sum splitting in general no longer is direct. Now we can choose the complementary spaces  $G^m$  such that

$$G^m \subset \ker \text{ad}_m A_s,$$

ensuring equivariance of the normalized part of  $\Phi_\star X$ , with respect to all linear transformations

$$\exp tA_s, \quad t \in \mathbb{R}.$$

The choice of  $G^m$  can be further restricted, e. g., such that

$$G^m \subseteq \ker \text{ad}_m A_s \setminus \text{im } \text{ad}_m A_n \subseteq H^m(\mathbb{R}^n),$$

compare with Van der Meer [88]. For further discussion on the choice of  $G^m$ , see below.

*Example (Rotational symmetry [84])* Consider the case  $n = 2$  where

$$A = A_s = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

So, the eigenvalues of  $A$  are  $\pm i$  and the transformations  $\exp tA_s$ ,  $t \in \mathbb{R}$ , form the rotation group  $\text{SO}(2, \mathbb{R})$ . From this, we can arrive at once at the general format of the normal form. In fact, the normalized,  $N$ th order part of  $\Phi_\star X$  is rotationally symmetric. This implies that, if we pass to polar coordinates  $(r, \varphi)$ , by

$$y_1 = r \cos \varphi, \quad y_2 = r \sin \varphi,$$

the normalized truncation of  $\Phi_\star X$  obtains the form

$$\begin{aligned}\dot{\varphi} &= f(r^2) \\ \dot{r} &= rg(r^2),\end{aligned}$$

for certain polynomials  $f$  and  $g$ , with  $f(0) = 1$  and  $g(0) = 0$ .

*Remark*

- A more direct, ‘computational’ proof of this result can be given as follows, compare with the proof of Theorem 2 and with [16,84,87]. Indeed, in vector field notation we can write

$$\begin{aligned}A &= -x_2 \frac{\partial}{\partial x_1} + x_1 \frac{\partial}{\partial x_2} \\ &= i \left( z \frac{\partial}{\partial z} - \bar{z} \frac{\partial}{\partial \bar{z}} \right),\end{aligned}$$

where we complexified putting  $z = x_1 + ix_2$  and use the well-known Wirtinger derivatives

$$\frac{\partial}{\partial z} = \frac{1}{2} \left( \frac{\partial}{\partial x_1} - i \frac{\partial}{\partial x_2} \right), \quad \frac{\partial}{\partial \bar{z}} = \frac{1}{2} \left( \frac{\partial}{\partial x_1} + i \frac{\partial}{\partial x_2} \right).$$

Now a basis of eigenvectors for  $\text{ad}_m A$  can be found directly, just computing a few Lie-brackets, compare the previous subsection. In fact, it is now given by all monomials

$$z^k \bar{z}^\ell \frac{\partial}{\partial z}, \quad \text{and} \quad z^k \bar{z}^\ell \frac{\partial}{\partial \bar{z}},$$

with  $k + \ell = m$ . The corresponding eigenvalues are  $i(k - \ell - 1)$  viz.  $i(k - \ell + 1)$ . So again we see, now by a direct inspection, that  $\text{ad}_m A$  is semisimple and that we can take  $G^m = \ker \text{ad}_m A$ . This space is spanned by

$$(z\bar{z})^r \left( z \frac{\partial}{\partial z} \pm \bar{z} \frac{\partial}{\partial \bar{z}} \right),$$

with  $2r + 1 = m - 1$ , which indeed proves that the normal form is rotationally symmetric.

- A completely similar case occurs for  $n = 3$ , where

$$A = A_s = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In this case the normalized part in cylindrical coordinates  $(r, \varphi, z)$ , given by

$$y_1 = r \cos \varphi, \quad y_2 = r \sin \varphi, \quad y_3 = z,$$

in general gets the axially symmetric form

$$\begin{aligned} \dot{\varphi} &= f(r^2, z) \\ \dot{r} &= rg(r^2, z) \\ \dot{z} &= h(r^2, z), \end{aligned}$$

for suitable polynomials  $f, g$  and  $h$ .

As a generalization of this example we state the following proposition, where the normalized part exhibits an  $m$ -torus symmetry.

**Proposition 7 (Toroidal symmetry [84])** *Let  $X$  be a  $C^\infty$  vector field, defined in the neighborhood of  $0 \in \mathbb{R}^n$ , with  $X(0) = 0$  and where  $A = D_0 X$  is semisimple with the eigenvalues  $\pm i\omega_1, \pm i\omega_2, \dots, \pm i\omega_m$  and  $0$ . Here  $2m \leq n$ . Suppose that for given  $N \in \mathbb{N}$  and all integer vectors  $(k_1, k_2, \dots, k_m)$ ,*

$$1 \leq \sum_{j=1}^m |k_j| \leq N + 1 \Rightarrow \sum_{j=1}^m k_j \omega_j \neq 0, \quad (1)$$

(i.e., there are no resonances up to order  $N + 1$ ). Then there exists, near  $0 \in \mathbb{R}^n$ , an analytic change of coordinates  $\Phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ , with  $\Phi(0) = 0$ , such that  $\Phi_* X$ , up to terms of order  $N$  has the following form. In suitable generalized cylindrical coordinates  $(\varphi_1, \dots, \varphi_m, r_1^2, \dots, r_m^2, z_{n-2m+1}, \dots, z_n)$  it is given by

$$\begin{aligned} \dot{\varphi}_j &= f_j(r_1^2, \dots, r_m^2, z_{n-2m+1}, \dots, z_n) \\ \dot{r}_j &= r_j g_j(r_1^2, \dots, r_m^2, z_{n-2m+1}, \dots, z_n) \\ \dot{z}_\ell &= h_\ell(r_1^2, \dots, r_m^2, z_{n-2m+1}, \dots, z_n), \end{aligned}$$

where  $f_j(0) = \omega_j$  and  $h_\ell(0) = 0$  for  $1 \leq j \leq m, n - 2m + 1 \leq \ell \leq n$ .

Proofs can be found, e.g., in [16,84,85,87]. In fact, if one introduces a suitable complexification, it runs along the same lines as the above remark. For the fact that finitely many non-resonance conditions are needed in order to normalize up to finite order, also compare a remark following Theorem 2.

Since the truncated system of  $\dot{r}_j$ - and  $\dot{z}_\ell$ -equations is independent of the angles  $\varphi_j$ , this can be studied separately. A similar remark holds for the earlier examples of this section. As indicated in the introduction, this kind of ‘reduction by symmetry’ to lower dimension can be of great importance when studying the dynamics of  $X$ : as it enables us to consider  $X$ , viz.  $\Phi_* X$ , as an  $N$ -flat perturbation of the normalized part, which is largely determined by this lower dimensional reduction. For more details see below.

### On the Choices of the Complementary Space and of the Normalizing Transformation

In the previous subsection we only provided a general (symmetric) format of the normalized part. In concrete examples one has to do more. Indeed, given the original Taylor series, one has to *compute* the coefficients in the normalized expansion. This means that many choices have to be made explicit.

To begin with there is the choice of the spaces  $G^m$ , which define the notion of ‘simple’. We have seen already that this choice is not unique.

Moreover observe that, even if the choice of  $G^m$  has been fixed, still  $P$  usually is not uniquely determined. In the semisimple case, for example,  $P$  is only determined modulo the kernel  $G^m = \ker \text{ad}_m A$ .

**Remark** To fix thoughts, consider the former of the above examples, on  $\mathbb{R}^2$ , where the normalized truncation has the

rotationally symmetric form

$$\begin{aligned}\dot{\varphi} &= f(r^2) \\ \dot{r} &= rg(r^2).\end{aligned}$$

Here  $g(r^2) = cr^2 + O(|r|^4)$ . The coefficient  $c$  dynamically is important, just think of the case where the system is part of a family that goes through Hopf-bifurcation. The computation of (the sign of)  $c$  in a concrete model can be quite involved, as it appears from, e.g., Marsden and McCracken [56]. Machine-assisted methods largely have taken over this kind of work.

One general way to choose  $G^m$  is the following, compare with e.g., Sect. 7.6 in [16] and Sect. 2.3 in [87]:

$$G^m := \ker(\text{ad}_m A^T).$$

Here  $A^T$  is the *transpose* of  $A$ , defined by the relation  $\langle A^T x, y \rangle = \langle x, Ay \rangle$ , where  $\langle \cdot, \cdot \rangle$  is an inner product on  $\mathbb{R}^n$ . A suitable choice for an inner product on  $H^m(\mathbb{R}^n)$  then directly gives that

$$G^m \oplus \text{im}(\text{ad}_m A) = H^m(\mathbb{R}^n),$$

as required. Also here the normal form can be interpreted in terms of symmetry, namely with respect to the group generated by  $A^T$ . In the semisimple case, this choice leads to exactly the same symmetry considerations as before.

The above algorithms do not provide methods for computing the normal form yet, i.e., for actually solving the homological equation. In practice, this is an additional computation. Regarding the corresponding algorithms we give a few more references for further reading, also referring to their bibliographies. General work in this direction is Bruno [35,36], Sect. 7.6 in Dumortier et al. [16], Takens [84] or Vanderbauwhede [87], Part 2. In the latter reference also a brief description is given of the  $\text{sl}(2, \mathbb{R})$ -theory of Sanders and Cushman [39], which is a powerful tool in the case where the matrix  $A$  is nilpotent [84]. For a thorough discussion on some of these methods we refer to Murdock [60]. Later on we shall come back to these aspects.

## Preservation of Structure

It goes without saying that Normal Form Theory is of great interest in special cases where a given structure has to be preserved. Here one may think of a symplectic or a volume form that has to be respected. Also a given symmetry group can have this rôle, e.g., think of an involution related to reversibility. Another, similar, problem is the dependence of external parameters in the system.

A natural language for preservation of such structures is that of Lie-subalgebra's of general Lie-algebra of vector fields, and the corresponding Lie-subgroup of the general Lie-group of diffeomorphisms.

## The Lie-Algebra Proof

Fortunately the setting of Theorem 6 is almost completely in terms of Lie-brackets. Let us briefly reconsider its proof.

Given is a  $C^\infty$  vector field  $X(x) = Ax + f(x)$ ,  $x \in \mathbb{R}^n$ , where  $A$  is linear and where  $f(0) = 0$ ,  $D_0 f = 0$ . We recall that in the inductive procedure a transformation

$$h = \text{Id} + P,$$

with  $P$  a homogeneous polynomial of degree  $m = 2, 3, \dots$ , is found, putting the homogeneous  $m$ th degree part of the Taylor series of  $X$  into the 'good' space  $G^m$ .

Now, nothing changes in this proof if instead of  $h = \text{Id} + P$ , we take  $h = P_1$ , the flow over time 1 of the vector field  $P$ : indeed, the effect of this change is not felt until the order  $2m - 1$ . Here we use the following formula for  $X^t := (P_t)_* X$ :

$$[X^t, P] = \text{ad}_m A(P) + O(|y|^{m+1})$$

$$\text{and } \frac{\partial X^t}{\partial t} = [X^t, P],$$

compare with Sect. "Preliminaries from Differential Geometry". In fact, exactly this choice  $h = P_1$  was taken by Roussarie [74], also see the proof of Takens [84]. Notice that  $\Phi = h_N \circ h_{N-1} \circ \dots \circ h_3 \circ h_2$ . Moreover notice that if the vector field  $P$  is in a given Lie-algebra of vector fields, its time 1 map  $P_1$  is in the corresponding Lie-group. In particular, if  $P$  is Hamiltonian,  $P_t$  is canonical or symplectic, and so on.

For the validity of this set-up for a more general Lie-subalgebra of the Lie-algebra of all  $C^\infty$  vector fields, one has to study how far the *grading*

$$\prod_{m=1}^{\infty} H^m(\mathbb{R}^n)$$

of the formal power series, as well as the *splittings*

$$B^m \oplus G^m = H^m(\mathbb{R}^n),$$

are compatible with the Lie-algebra at hand. This issue was addressed by Broer [7,8] axiomatically, in terms of graded and filtered Lie-algebra's. Moreover, the methods concerning the choice of  $G^m$  briefly mentioned at the end of the previous section, all carry over to a more general Lie-algebra set-up. As a consequence there exists a version of Theorem 6 in this setting. Instead of pursuing this further, we discuss its implications in a few relevant settings.



**The Volume Preserving and Symplectic Case** On  $\mathbb{R}^n$ , resp.  $\mathbb{R}^{2n}$ , we consider a volume form or a symplectic form, both denoted by  $\sigma$ . We assume, that

$$\sigma = dx_1 \wedge \cdots \wedge dx_n, \quad \text{resp.} \quad \sigma = \sum_{j=1}^n dx_j \wedge dx_{j+n}.$$

In both cases, let  $\mathcal{X}_\sigma$  denote the Lie-algebra of  $\sigma$ -preserving vector fields, i.e., vector fields  $X$  such that  $\mathcal{L}_X \sigma = 0$ . Here  $\mathcal{L}$  again denotes the Lie-derivative, see Sect. “[The Normal Form Procedure](#)”.

Indeed, one defines

$$\begin{aligned} \mathcal{L}_X \sigma(x) &= \left. \frac{d}{dt} \right|_{t=0} (X_t)^* \sigma(x) \\ &= \lim_{t \rightarrow 0} \frac{1}{t} \{ (X_t)^* \sigma(x) - \sigma(x) \}. \end{aligned}$$

Properties, similar to Proposition 5, hold here. Since in both cases  $\sigma$  is a closed form, one shows by ‘the magic formula’ [81] that

$$\begin{aligned} \mathcal{L}_X \sigma(x) &= d(\iota_X \sigma) + \iota_X d\sigma \\ &= d(\iota_X \sigma). \end{aligned}$$

Here  $\iota_X \sigma$  denotes the flux-operator defined by  $\iota_X \sigma(Y) = \sigma(X, Y)$ . In the volume-preserving case the latter expression denotes  $\text{div}(X)\sigma$  and we see that preservation of  $\sigma$  exactly means that  $\text{div}(X) = 0$ : the divergence of  $X$  vanishes. In the Hamiltonian case we conclude that the 1-form  $\iota_X \sigma$  is closed and hence (locally) of the form  $dH$ , for a Hamilton function  $H$ . In both cases, the fact that for a transformation  $h$  the fact that  $h^* \sigma = \sigma$  implies that with  $X$  also  $h_* X$  is  $\sigma$ -preserving. Moreover, for a  $\sigma$ -preserving vector field  $P$  and  $h = P_1$  one can show that indeed  $h^* \sigma = \sigma$ .

One other observation is, that by the *homogeneity* of the above expressions for  $\sigma$ , the homogeneous parts of the Taylor series of  $\sigma$ -preserving vector fields are again  $\sigma$ -preserving. This exactly means that

$$H^m(\mathbb{R}^{(2)n}) \cap \mathcal{X}_\sigma,$$

$m = 1, 2, \dots$ , grades the formal power series corresponding to  $\mathcal{X}_\sigma$ . Here, notice that

$$H^1(\mathbb{R}^{(2)n}) \cap \mathcal{X}_\sigma = \mathfrak{sl}(n, \mathbb{R}), \quad \text{resp.} \quad \mathfrak{sp}(2n, \mathbb{R}),$$

the *special*- resp. the *symplectic* linear algebra.

In summary we conclude that both the symplectic and the volume preserving setting are covered by the axiomatic approach of [7,8] and that an appropriate version of Theorem 6 holds here. Below we shall illustrate this with a few examples.

**External Parameters** A  $C^\infty$  family  $X = X^\lambda(x)$  of vector fields on  $\mathbb{R}^n$ , with a multi-parameter  $\lambda \in \mathbb{R}^p$ , can be regarded as one  $C^\infty$  vector field on the product space  $\mathbb{R}^n \times \mathbb{R}^p$ . Such a vector field is *vertical*, in the sense that it has no components in the  $\lambda$ -direction. In other words, if  $\pi : \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}^p$  is the natural projection on the second component,  $X$  is tangent to the fibers of  $\pi$ . It is easily seen that this property defines a Lie-subalgebra of the Lie-algebra of all  $C^\infty$  vector fields on  $\mathbb{R}^n \times \mathbb{R}^p$ . Again, by the linearity of this projection, the gradings and splittings are compatible. The normal form transformations  $\Phi$  preserve the parameter  $\lambda$ , i.e.,  $\Phi \circ \pi = \pi$ .

When studying a bifurcation problem, we often consider systems  $X = X^\lambda(x)$  locally defined near  $(x, \lambda) = (0, 0)$ , considering series expansions both in  $x$  and in  $\lambda$ . Then, in the  $N$ th order normalization  $\Phi_* X$ , the normalized part consists of a polynomial in  $y$  and  $\lambda$ , while the remainder term is of the form  $O(|y|^{N+1} + |\lambda|^{N+1})$ .

As in the previous case, we shall not formulate the present analogue of Theorem 6 for this case, but illustrate its meaning in examples.

**The Reversible Case** In the reversible case a linear involution  $R$  is given, while for the vector fields we require  $R_* X = -X$ . Let  $\mathcal{X}_R$  denote the class of all such reversible vector fields. Also, let  $\mathcal{C}$  denote the class of all  $X$  such that  $R_* X = X$ . Then, both  $\mathcal{X}_R$  and  $\mathcal{C}$  are linear spaces of vector fields. Moreover,  $\mathcal{C}$  is a Lie-subalgebra. Associated to  $\mathcal{C}$  is the group of diffeomorphisms that commute with  $R$ , i.e., the *R-equivariant* transformations. Also it is easy to see that for each of these diffeomorphisms  $\Phi$  one has  $\Phi_*(\mathcal{X}_R) \subset \mathcal{X}_R$ . The above approach applies to this situation in a straightforward manner. The gradings and splittings fit, while we have to choose the infinitesimal generator  $P$  from the set  $\mathcal{C}$ . For details compare with [29].

**Remark** In the case with parameters, it sometimes is possible to obtain an alternative normal form where the normalized part is polynomial in  $y$  alone, with coefficients that depend smoothly on  $\lambda$ . A necessary condition for this is that the origin  $y = 0$  is an equilibrium for all values of  $\lambda$  in some neighborhood of 0. To be precise, at the  $N$ th order we can achieve smooth dependence of the coefficients on  $\lambda$  for  $\lambda \in \Lambda_N$ , where  $\Lambda_N$  is a neighborhood of  $\lambda = 0$ , that may shrink to  $\{0\}$  as  $N \rightarrow \infty$ . So, for  $N \rightarrow \infty$  only the formal aspect remains, as is the case in the above approach. This alternative normal form can be obtained by a proper use of the Implicit Function Theorem in the spaces  $H^m(\mathbb{R}^n)$ ; for details e.g., see Section 2.2 in Vanderbauwhede [87]. For another discussion on this topic, cf. Section 7.6.2 in Dumortier et al. [16].

In Sect. “[The Normal Form Procedure](#)” the role of symmetry was considered regarding the semisimple part of the matrix  $A$ . A question is how this discussion generalizes to Lie-subalgebra’s of vector fields.

*Example (Volume preserving, parameter dependent axial symmetry [7,8])* On  $\mathbb{R}^3$  consider a 1-parameter family  $X^\lambda$  of vector fields, preserving the standard volume  $\sigma = dx_1 \wedge dx_2 \wedge dx_3$ . Assume that  $X^0(0) = 0$  while the spectrum of  $D_0 X^0$  consists of the eigenvalues  $\pm i$  and 0. For the moment regarding  $\lambda$  as an extra state space coordinate, we obtain a vertical vector field on  $\mathbb{R}^4$  and we apply a combination of the above considerations. The ‘generic’ Jordan normal form then is

$$A = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (2)$$

with an obvious splitting in semisimple and nilpotent part. The considerations of Sect. “[The Normal Form Procedure](#)” then directly apply to this situation. For any  $N$  this yields a transformation  $\Phi : \mathbb{R}^4 \rightarrow \mathbb{R}^4$ , with  $\Phi(0) = 0$ , preserving both the projection to the 1-dimensional parameter space and the volume of the 3-dimensional phase space, such that the normalized,  $N$ th order part of  $\Phi_* X(y, \lambda)$ , in cylindrical coordinates  $y_1 = r \cos \varphi$ ,  $y_2 = r \sin \varphi$ ,  $y_3 = z$ , has the rotationally symmetric form

$$\begin{aligned} \dot{\varphi} &= f(r^2, z, \lambda) \\ \dot{r} &= rg(r^2, z, \lambda) \\ \dot{z} &= h(r^2, z, \lambda), \end{aligned}$$

again, for suitable polynomials  $f, g$  and  $h$ . Note, that in cylindrical coordinates the volume has the form  $\sigma = r dr \wedge d\varphi \wedge dz$ . Again the functions  $f, g$  and  $h$  have to fit with the linear part. In particular we find that  $h(r^2, z, \lambda) = \lambda + az + \dots$ , observing that for  $\lambda \neq 0$  the origin is no equilibrium point.

*Remark* If  $A = A_s + A_n$  is the canonical splitting of  $A$  in  $H^1(\mathbb{R}^n) = \mathfrak{gl}(n, \mathbb{R})$ , then automatically both  $A_s$  and  $A_n$  are in the subalgebra under consideration. In the volume preserving setting this can be seen directly. In general the same holds true as soon as the corresponding linear Lie-group is algebraic, see [7,8] and the references given there.

### The Hamiltonian Case

The Normal Form Theory in the Hamiltonian case goes back at least to Poincaré [69] and Birkhoff [5]. Other references are, for instance, Gustavson [47], Arnold [1,2,3],

Sanders, Verhulst and Murdock [75], Van der Meer [88], Broer, Chow and Kim [17]. In Sect. “[Preservation of Structure](#)” we already saw that the axiomatic Lie algebra approach of [7,8] applies here, especially since the symplectic group  $\mathrm{SP}(2n, \mathbb{R})$  is algebraic. The canonical form here usually goes with the name Williamson, compare Galin [44], Koçak [54] and Hoveijn [49]. We discuss how the Lie algebra approach compares to the literature.

The Lie algebra of Hamiltonian vector fields can be associated to the Poisson-algebra of Hamilton functions as follows, even in an arbitrary symplectic setting. As before, the symplectic form is denoted by  $\sigma$ . We recall, that for any Hamiltonian  $H$  the corresponding Hamiltonian vector field  $X_H$  is given by  $dH = \sigma(X_H, \cdot)$ . Now, let  $H$  and  $K$  be Hamilton functions with corresponding vector fields  $X_H$  resp.  $X_K$ . Then

$$X_{\{H,K\}} = [X_H, X_K],$$

implying that the map  $H \mapsto X_H$  is a morphism of Lie algebra’s. By definition this map is surjective, while its kernel consists of the (locally) constant functions.

This implies, that the normal form procedure can be completely rephrased in terms of the Poisson-bracket. We shall now demonstrate this by an example, similar to the previous one.

*Example (Symplectic parameter dependent rotational symmetry [17])* Consider  $\mathbb{R}^4$  with coordinates  $(x_1, y_1, x_2, y_2)$  and the standard symplectic form  $\sigma = dx_1 \wedge dy_1 + dx_2 \wedge dy_2$ , considering a  $C^\infty$  family of Hamiltonian functions  $H^\lambda$ , where  $\lambda \in \mathbb{R}$  is a parameter. The fact that  $dH = \sigma(X_H, \cdot)$  in coordinates means

$$\dot{x}_j = \frac{\partial}{\partial y_j} H, \quad \dot{y}_j = -\frac{\partial}{\partial x_j} H,$$

for  $j = 1, 2$ . We assume that for  $\lambda = 0$  the origin of  $\mathbb{R}^4$  is a singularity. Then we expand as a Taylor series in  $(x, y, \lambda)$

$$H^\lambda(x, y) = H_2(x, y, \lambda) + H_3(x, y, \lambda) + \dots,$$

where the  $H_m$  is homogeneous of degree  $m$  in  $(x, y, \lambda)$ . It follows for the corresponding Hamiltonian vector fields that

$$X_{H_m} \in H^{m-1}(\mathbb{R}^4),$$

in particular  $X_{H_2} \in \mathfrak{sp}(4, \mathbb{R}) \subset H^1(\mathbb{R}^4)$ . Let us assume that this linear part  $X_{H_2}$  at  $(x, y, \lambda) = (0, 0, 0)$  has eigenvalues  $\pm i$  and a double eigenvalue 0. One ‘generic’ Williamson’s normal form then is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad (3)$$

compare with (2), corresponding to the quadratic Hamilton function

$$H_2(x, y, \lambda) = I + \frac{1}{2}y_2^2,$$

where  $I := \frac{1}{2}(x_1^2 + y_1^2)$ . It is straightforward to give the generic matrix for the linear part in the extended state space  $\mathbb{R}^5$ , see above, so we will leave this to the reader.

The semisimple part  $A_s = X_I$  now can be used to obtain a rotationally symmetric normal form, as before. In fact, for any  $N \in \mathbb{N}$  there exists a canonical transformation  $\Phi(x, y, \lambda)$ , which keeps the parameter fixed, and a polynomial  $F(I, x_2, y_2, \lambda)$ , such that

$$(H \circ \Phi^{-1})(x, y, \lambda) = F(I, x_2, y_2, \lambda) + O(I + x_2^2 + y_2^2 + \lambda^2)^{(N+1)/2}.$$

Instead of using the adjoint action  $\text{ad}_A$  on the spaces  $H^{m-1}(\mathbb{R}^5)$ , we also may use the adjoint action

$$\text{ad}H_2 : f \mapsto \{H_2, f\},$$

where  $f$  is a polynomial function of degree  $m$  in  $(x_1, y_1, x_2, y_2, \lambda)$ . In the vector field language, we choose the ‘good’ space  $G^{m-1} \subset \ker \text{ad}_{m-1}A_s$ , which, in the function language, translates to a ‘good’ subset of  $\ker \text{ad}I$ .

Whatsoever, the normalized part of the Hamilton function  $H \circ \Phi^{-1}$ , viz. the vector field  $\Phi_* X_H = X_{H \circ \Phi^{-1}}$ , is rotationally symmetric. The fact that the Hamilton function  $F$  Poisson-commutes with  $I$  exactly amounts to invariance under the action generated by the vector field  $X_I$ , in turn implying that  $I$  is an integral of  $X_F$ . Indeed, if we define a  $2\pi$ -periodic variable  $\varphi$  as follows:

$$x_1 = \sqrt{2I} \sin \varphi, \quad y_1 = \sqrt{2I} \cos \varphi,$$

then  $\sigma = dI \wedge d\varphi + dx_2 \wedge dy_2$ , implying that the normalized vector field  $X_F$  has the canonical form

$$\begin{aligned} \dot{I} &= 0, \quad \dot{\varphi} = -\frac{\partial F}{\partial I}, \\ \dot{x}_2 &= \frac{\partial F}{\partial y_2}, \quad \dot{y}_2 = -\frac{\partial F}{\partial x_2}. \end{aligned}$$

Notice that  $\varphi$  is a cyclic variable, making the fact that  $I$  is an integral clearly visible. Also observe that  $\dot{\varphi} = -1 + \dots$ . As before, and as in, e.g., the central force problem, this enables a reduction to lower dimension. Here, the latter two equations constitute the *reduction to 1 degree of freedom*: it is a family of planar Hamiltonian vector fields, parametrized by  $I$  and  $\lambda$ .

*Remark*

- This example has many variations. First of all it can be simplified by omitting parameters and even the zero eigenvalues. The conclusion then is that a planar Hamilton function with a nondegenerate minimum or maximum has a formal rotational symmetry, up to canonical coordinate changes.
- It also can be easily made more complicated, compare with Proposition 7 in Sect. “[The Normal Form Procedure](#)”. Given an equilibrium with purely imaginary eigenvalues  $\pm i\omega_1, \pm i\omega_2, \dots, \pm i\omega_m$  and with  $2(n-m)$  zero eigenvalues. Provided that there are no resonances up to order  $N+1$ , see (1), also here we conclude that Hamiltonian truncation at the order  $N$ , by canonical transformations can be given the form

$$F(I_1, \dots, I_m, x_{m+1}, y_{m+1}, \dots, x_n, y_n),$$

where  $I_j := \frac{1}{2}(x_j^2 + y_j^2)$ . As in Proposition 7 this normal form has a toroidal symmetry. Writing  $x_j = \sqrt{2I_j} \sin \varphi_j$ ,  $y_j = \sqrt{2I_j} \cos \varphi_j$ , we obtain the canonical system of equations

$$\begin{aligned} \dot{I}_j &= 0, \quad \dot{\varphi}_j = -\frac{\partial F}{\partial I_j}, \\ \dot{x}_\ell &= \frac{\partial F}{\partial y_\ell}, \quad \dot{y}_\ell = -\frac{\partial F}{\partial x_\ell}, \end{aligned}$$

$1 \leq j \leq m$ ,  $m+1 \leq \ell \leq n$ . Note that  $\dot{\varphi}_j = -\omega_j + \dots$ . In the case  $m = n$  we deal with an elliptic equilibrium. The corresponding result usually is named the Birkhoff normal form [1,5,47], ► [Dynamics of Hamiltonian Systems](#). Then, the variables  $(I_j, \varphi_j)$ ,  $1 \leq j \leq m$ , are a set of *action-angle variables* for the truncated part of order  $N$ , [1].

- Returning to algorithmic issues, in addition to Subsect. “[On the Choices of the Complementary Space and of the Normalizing Transformation](#)”, we give a few further references in cases where a structure is being preserved: Broer [7], Deprit [40,41], Meyer [57] and Ferrer, Hanßmann, Palacián and Yanguas [43]. It is to be noted that this kind of Hamiltonian result also can be obtained, where the coordinate changes are constructed using *generating functions*, compare with, e.g., [1,3,77], see also the discussion in Broer, Hoveijn, Lunter and Vegter [20] and in particular Section 4.4.2 in [23]. For another, computationally very effective normal form algorithm, see Giorgilli and Gallgani [45]. Also compare with references therein.

## Semi-local Normalization

This section roughly consists of two parts. To begin with, a number of subsections are devoted to related formal normal form results, near fixed points of diffeomorphisms and near periodic solutions and invariant tori of vector fields.

### A Diffeomorphism Near a Fixed Point

We start by formulating a result by Takens:

**Theorem 8 (Takens normal form [83])** *Let  $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a  $C^\infty$  diffeomorphism with  $T(0) = 0$  and with a canonical decomposition of the derivative  $D_0T = S + N$  in semisimple resp. nilpotent part. Also, let  $N \in \mathbb{N}$  be given, then there exists diffeomorphism  $\Phi$  and a vector field  $X$ , both of class  $C^\infty$ , such that  $S_*X = X$  and*

$$\Phi^{-1} \circ T \circ \Phi = S \circ X_1 + O(|y|^{N+1}).$$

Here, as before,  $X_1$  denotes the flow over time 1 of the vector field  $X$ . Observe that the vector field  $X$  necessarily has the origin as an equilibrium point. Moreover, since  $S_*X = X$ , the vector field  $X$  is invariant with respect to the group generated by  $S$ .

The proof is a bit more involved than Theorem 6 in Sect. “The Normal Form Procedure”, but it has the same spirit, also compare with [37]. In fact, the Taylor series of  $T$  is modified step-by-step, using coordinate changes generated by homogeneous vector fields of the proper degree.

After a reduction to center manifolds the spectrum of  $S$  is on the complex unit circle and Theorem 8 especially is of interest in cases where this spectrum consists of roots of unity, i. e., in the case of resonance. Compare with [32,37,83].

Again, the result is completely phrased in terms of Lie algebra's and groups and therefore bears generalization to many contexts with a preserved structure, compare Sect. “Preservation of Structure”. The normalizing transformations of the induction process then are generated from the corresponding Lie algebra. For a symplectic analogue see Moser [59]. Also, both in Broer, Chow and Kim [17] and in Broer and Vegter [14], symplectic cases with parameters are discussed, where  $S = \text{Id}$ , the Identity Map, resp.  $S = -\text{Id}$ , the latter involving a period-doubling bifurcation.

**Remark** Let us consider a symplectic map  $T$  of the plane, which means that  $T$  preserves both area and orientation. Assume that  $T$  is fixing the origin, while the eigenvalues of  $S = D_0T$  are on the unit circle, without being roots of unity. Then  $S$  generates the rotation group  $\text{SO}(2, \mathbb{R})$ , so the vector field  $X$ , which has divergence zero, in this case

is rotationally symmetric. Again this result often goes with the name of Birkhoff.

### Near a Periodic Solution

The Normal Form Theory at a periodic solution or closed orbit has a lot of resemblance to the local theory we met before.

To fix thoughts, let us consider a  $C^\infty$  vector field of the form

$$\begin{aligned}\dot{x} &= f(x, y) \\ \dot{y} &= g(x, y),\end{aligned}\tag{4}$$

with  $(x, y) \in \mathbb{T}^1 \times \mathbb{R}^n$ . Here  $\mathbb{T}^1 = \mathbb{R}/(2\pi\mathbb{Z})$ . Assuming  $y = 0$  to be a closed orbit, we consider the formal Taylor series with respect to  $y$ , with  $x$ -periodic coefficients. By Floquet Theory, we can assume that the coordinates  $(x, y)$  are such that

$$f(x, y) = \omega + O(|y|), \quad g(x, y) = \Omega y + O(|y|^2),$$

where  $\omega \in \mathbb{R}$  is the frequency of the closed orbit and  $\Omega \in \text{gl}(n, \mathbb{R})$  its Floquet matrix. Again, the idea is to ‘simplify’ this series further. To this purpose we introduce a grading as before, letting  $H^m = H^m(\mathbb{T}^1 \times \mathbb{R}^n)$  be the space of vector fields

$$Y(x, y) = L(x, y) \frac{\partial}{\partial x} + \sum_{j=1}^n M_j(x, y) \frac{\partial}{\partial y_j},$$

with  $L(x, y)$  and  $M_j(x, y)$  homogeneous in  $y$  of degree  $m-1$  resp.  $m$ . Notice, that this space  $H^m$  is infinite-dimensional. However, this is not at all problematic for the things we are doing here. By this definition, we have that

$$A := \omega \frac{\partial}{\partial x} + \Omega y \frac{\partial}{\partial y}$$

is a member of  $H^1$  and with this normally linear part we can define an adjoint representation  $\text{ad}A$  as before, together with linear maps

$$\text{ad}_m A : H^m \rightarrow H^m.$$

Again we assume to have a decomposition

$$G^m \oplus \text{Im}(\text{ad}_m A) = H^m,$$

where the aim is to transform the terms of the series successively into the  $G^m$ , for  $m = 2, 3, 4, \dots$

The story now runs as before. In fact, the proof of Theorem 6 in Sect. “The Normal Form Procedure”, as well as

its Lie algebra versions indicated in Sect. “[Preservation of Structure](#)”, can be repeated almost verbatim for this case. Moreover, if  $\Omega = \Omega_s + \Omega_n$  is the canonical splitting in semisimple and nilpotent part, then

$$\omega \frac{\partial}{\partial x} + \Omega_s y \frac{\partial}{\partial y}$$

gives the semisimple part of  $\text{ad}_m A$ , as can be checked by a direct computation. From this computation one also deduces the non-resonance conditions needed for the present torus-symmetric analogue of Proposition 7 in Sect. “[The Normal Form Procedure](#)”.

There are different cases *with* resonance either between the imaginary parts of the eigenvalues of  $\Omega$  (normal resonance) or between the latter and the frequency  $\omega$  (normal-internal resonance). All of this extends to the various settings with preservation of structure as discussed before. In all cases direct analogues of the Theorems 6 and 8 are valid. General references in this direction are Arnold [2,3], Bruno [35,36], Chow, Li and Wang [37], Iooss [50], Murdock [60] or Sanders, Verhulst and Murdock [75].

#### Remark

- This approach also is important for non-autonomous systems with periodic time dependence. Here the normalization procedure includes averaging. As a special case of the above form, we obtain a system

$$\begin{aligned}\dot{x} &= \omega \\ \dot{y} &= g(x, y),\end{aligned}$$

so where  $x \in \mathbb{T}^1$  is proportional to the time. Apart from the general references given above, we also refer to, e.g., Broer and Vegter [14] and to Broer, Roussarie and Simó [9,19]. The latter two applications also contain parameters and deal with bifurcations. Also compare with Verhulst ► [Perturbation Analysis of Parametric Resonance](#).

- A geometric tool that can be successfully applied in various resonant cases is a *covering space*, obtained by a Van der Pol transformation (or by passing to co-rotating coordinates). This involves equivariance with respect to the corresponding deck group. This setting (with or without preservation of structure) is completely covered by the general Lie algebra approach as described above.

For the Poincaré map this deck group symmetry directly yields the normal form symmetry of Theorem 8. For applications in various settings, with or without preservation of structure, see [14,24,32]. This normalization technique is effective for studying bifurcation of

subharmonic solutions. In the case of period doubling the covering space is just a double cover. In many cases Singularity Theory turns out to be useful.

#### Near a Quasi-periodic Torus

The approach of the previous subsection also applies at an invariant torus, provided that certain requirements are met. Here we refer to Braaksma, Broer and Huitema [15], Broer and Takens [12] and Bruno [35,36].

Let us consider a  $C^\infty$ -system

$$\begin{aligned}\dot{x} &= f(x, y) \\ \dot{y} &= g(x, y)\end{aligned}\tag{5}$$

as before, with  $(x, y) \in \mathbb{T}^m \times \mathbb{R}^n$ . Here  $\mathbb{T}^m = \mathbb{R}^m / (2\pi\mathbb{Z})^m$ . We assume that  $f(x, y) = \omega + O(|y|)$ , which implies that  $y = 0$  is an invariant  $m$ -torus, with on it a constant vector field with frequency-vector  $\omega$ . We also assume that  $g(x, y) = \Omega y + O(|y|^2)$ , which is the present analogue of the Floquet form as known in the periodic case, with  $\Omega \in \text{gl}(n, \mathbb{R})$ , independent of  $x$ . Contrary to the situations for  $m = 1$  and  $n = 1$ , in general reducibility to Floquet form is not possible. Compare with [10,15,18,38] and references therein. For a similar approach of a system that is not reducible to Floquet form, compare with [22].

Presently we assume this reducibility, expanding in formal series with respect to the variables  $y$ , where the coefficients are functions on  $\mathbb{T}^m$ . These coefficients, in turn, can be expanded in Fourier series. The aim then is, to ‘simplify’ this combined series by successive coordinate changes, following the above procedure. As a second requirement it then is needed that certain *Diophantine* conditions are satisfied on the pair  $(\omega, \Omega)$ . Below we give more details on this. Instead of giving general results we again refer to [12,15,26,29,35,36]. Moreover, to fix thoughts, we present a simple example with a parameter. Here again a direct link with averaging holds.

*Example (Toroidal symmetry with small divisors [38])*

Given is a family of vector fields  $\dot{x} = X(x, \lambda)$  with  $(x, \lambda) \in \mathbb{T}^m \times \mathbb{R}$ . We assume that  $X = X(x, \lambda)$  has the form

$$X(x, \lambda) = \omega + f(x, \lambda),$$

with  $f(x, 0) \equiv 0$ . It is assumed that the frequency vector  $\omega = (\omega_1, \omega_2, \dots, \omega_n)$  has components that satisfy the Diophantine non-resonance condition

$$|\langle \omega, k \rangle| \geq \gamma |k|^{-\tau},\tag{6}$$



for all  $k \in \mathbb{Z} \setminus \{0\}$ . Here  $\gamma > 0$  and  $\tau > n - 1$  are prescribed constants. We note that  $\tau > n - 1$  implies that this condition in  $\mathbb{R}^n = \{\omega\}$  excludes a set of Lebesgue measure  $O(\gamma)$  as  $\gamma \downarrow 0$ , compare with [38]. It follows, that by successive transformations of the form

$$h: (x, \lambda) \mapsto (x + P(x, \lambda), \lambda)$$

the  $x$ -dependence of  $X$  can be pushed away to ever higher order in  $\lambda$ , leading to a formal normal form

$$\dot{\xi} = \omega + g(\lambda),$$

with  $g(0) = 0$ . Observe that in this case ‘simple’ means  $x$ - (or  $\xi$ -) independent. Therefore, in a proper formalism,  $x$ -independent systems constitute the spaces  $G^m$ . Indeed, in the induction process we get

$$\begin{aligned} X(x, \lambda) &= \omega + g_2(\lambda) + \cdots + g_{N-1}(\lambda) \\ &\quad + f_N(x, \lambda) + O(|\lambda|^{N+1}), \end{aligned}$$

compare the proof of Theorem 6 in Sect. “[The Normal Form Procedure](#)”. Writing  $\xi = x + P(x, \lambda)$ , with  $P(x, \lambda) = \tilde{P}(x, \lambda)\lambda^N$ , we substitute

$$\begin{aligned} \dot{\xi} &= (\text{Id} + D_\xi P)\dot{x} \\ &= \omega + g_2(\lambda) + \cdots + g_{N-1}(\lambda) \\ &\quad + f_N(x, \lambda) + D_x P(x, \lambda)\omega + O(|\lambda|^{N+1}), \end{aligned}$$

Where we express the right-hand side in  $x$ . So we have to satisfy an equation

$$D_x P(x, \lambda)\omega + f_N(x, \lambda) \equiv c\lambda^N \bmod O(|\lambda|^{N+1}),$$

for a suitable constant  $c$ . Writing  $f_N(x, \lambda) = \tilde{f}_N(x, \lambda)\lambda^N$ , this amounts to

$$D_x \tilde{P}(x, \lambda)\omega = -\tilde{f}_N(x, \lambda) + c,$$

which is the present form of the homological equation. If

$$\tilde{f}_N(x, \lambda) = \sum_{k \in \mathbb{Z}^n} a_k(\lambda) e^{i\langle x, k \rangle},$$

then  $c = a_0$ , i.e., the  $m$ -torus average

$$a_0(\lambda) = \frac{1}{(2\pi)^m} \int_{\mathbb{T}^m} \tilde{f}_N(x, \lambda) dx.$$

Moreover,

$$\tilde{P}(x, \lambda) = \sum_{k \neq 0} \frac{a_k(\lambda)}{i\langle \omega, k \rangle} e^{i\langle x, k \rangle}.$$

This procedure formally only makes sense if the frequencies  $(\omega_1, \omega_2, \dots, \omega_m)$  have no resonances, which means that for  $k \neq 0$  also  $\langle \omega, k \rangle \neq 0$ . In other words this means that the components of the frequency vector  $\omega$  are rationally independent. Even then, the denominator  $i\langle \omega, k \rangle$  can become arbitrarily small, so casting doubt on the convergence. This problem of *small divisors* is resolved by the Diophantine conditions (6). For further reference, e.g., see [2,10,15,18,38]: for real analytic  $X$ , by the Paley–Wiener estimate on the exponential decay of the Fourier coefficients, the solution  $P$  again is real analytic. Also in the  $C^\infty$ -case the situation is rather simple, since then the coefficients in both cases decay faster than any polynomial.

*Remark*

- The discussion at the end of Subsect. “[Near a Periodic Solution](#)” concerning normalization at a periodic solution, largely extends to the present case of a quasi-periodic torus. Assuming reducibility to Floquet form, we now have a normally linear part

$$\omega \frac{\partial}{\partial x} + \Omega y \frac{\partial}{\partial y},$$

with a frequency vector  $\omega \in \mathbb{R}^n$ . As before  $\Omega \in \text{gl}(m, \mathbb{R})$ . For the corresponding KAM Perturbation Theory the Diophantine conditions (6) on the frequencies are extended by including the normal frequencies, i.e., the imaginary parts  $\omega_1^N, \dots, \omega_s^N$  of the eigenvalues of  $\Omega$ . To be precise, for  $\gamma > 0$  and  $\tau > n - 1$ , above the conditions (6), these extra Melnikov conditions are given by

$$\begin{aligned} |\langle \omega, k \rangle + \omega_j^N| &\geq \gamma |k|^{-\tau} \\ |\langle \omega, k \rangle + 2\omega_j^N| &\geq \gamma |k|^{-\tau} \\ |\langle \omega, k \rangle + \omega_j^N + \omega_\ell^N| &\geq \gamma |k|^{-\tau}, \end{aligned} \tag{7}$$

for all  $k \in \mathbb{Z}^n \setminus \{0\}$  and for  $j, \ell = 1, 2, \dots, s$  with  $\ell \neq j$ . See below for the description of an application to KAM Theory (and more references) in the Hamiltonian case.

- In this setting normal resonances can occur between the  $\omega_j^N$  and normal-internal resonances between the  $\omega_j^N$  and  $\omega$ . Certain strong normal-internal resonances occur when one of the left-hand sides of (7) vanishes. For an example see below. Apart from these now also internal resonances between the components of  $\omega$  come into play. The latter generally lead to destruction of the invariant torus.
- As in the periodic case also here covering spaces turn out to be useful for studying various resonant

bifurcation scenarios often involving applications of both Singularity Theory and KAM Theory. Compare with [15,25,31,33].

### Non-formal Aspects

Up to this moment (almost) all considerations have been formal, i. e., in terms of formal power series. In general, the Taylor series of a  $C^\infty$ -function, say,  $\mathbb{R}^n \rightarrow \mathbb{R}$  will be divergent. On the other hand, any formal power series in  $n$  variables occurs as the Taylor series of some  $C^\infty$ -function. This is the content of a theorem by É. Borel, cf. Narasimhan [61].

We briefly discuss a few aspects regarding convergence or divergence of the normalizing transformation or of the normalized series. We recall that the growth rate of the formal series, including the convergent case, is described by the Gevrey index, compare with, e. g., [4,55,72,73].

### Normal Form Symmetry and Genericity

For the moment we assume that all systems are of class  $C^\infty$ . As we have seen, if the normalization procedure is carried out to some finite order  $N$ , the transformation  $\Phi$  is a real analytic map. If we take the limit for  $N \rightarrow \infty$ , we only get formal power series  $\hat{\Phi}$ , but, by the Borel Theorem, a ‘real’  $C^\infty$ -map  $\Phi$  exists with  $\hat{\Phi}$  as its Taylor series.

Let us discuss the consequences of this, say, in the case of Proposition 7 in Sect. “The Normal Form Procedure”. Assuming that there are no resonances at all between the  $\omega_j$ , as a corollary, we find a  $C^\infty$ -map  $y = \Phi(x)$  and a  $C^\infty$  vector field  $p = p(y)$ , such that:

- The vector field  $\Phi_*X - p$ , in corresponding generalized cylindrical coordinates has the symmetric  $C^\infty$ -form

$$\begin{aligned}\dot{\phi}_j &= f_j(r_1^2, \dots, r_m^2, z_{n-2m+1}, \dots, z_n) \\ \dot{r}_j &= r_j g_j(r_1^2, \dots, r_m^2, z_{n-2m+1}, \dots, z_n) \\ \dot{z}_\ell &= h_\ell(r_1^2, \dots, r_m^2, z_{n-2m+1}, \dots, z_n),\end{aligned}$$

where  $f_j(0) = \omega_j$  and  $h_\ell(0) = 0$  for  $1 \leq j \leq m$ ,  $n - 2m + 1 \leq \ell \leq n$ .

- The Taylor series of  $p$  identically vanishes at  $y = 0$ .

Note, that an  $\infty$ -ly flat term  $p$  can have component functions like  $e^{-1/y_1^2}$ . We see, that the  $m$ -torus symmetry only holds up to such flat terms. Therefore, this symmetry, if present at all, also can be destroyed again by a *generic* flat ‘perturbation’. We refer to Broer and Takens [12], and references therein, for further consequences of this idea. The

main point is, that by a Kupka–Smale argument, which generically forbids so much symmetry, compare with [67].

*Remark* The Borel Theorem also can be used in the reversible, the Hamiltonian and the volume preserving setting. In the latter two cases we exploit the fact that a structure preserving vector field is generated by a function, resp. an  $(n - 2)$ -form. Similarly the structure preserving transformations have such a generator. On these generators we then apply the Borel Theorem. Many Lie algebra’s of vector fields have this ‘Borel Property’, saying that a formal power series of a transformation can be represented by a  $C^\infty$  map in the same structure preserving setting.

### On Convergence

The above topological ideas also can be pursued in many real analytic cases, where they imply a *generic divergence* of the normalizing transformation. For an example in the case of the Hamiltonian Birkhof normal form [5,47,77] compare with Broer and Tangerman [13] and its references. As an example we now deal with the linearization of a holomorphic germ in the spirit of Sect. “The Normal Form Procedure”.

*Example (Holomorphic linearization [34,58,92])* A holomorphic case concerns the linearization of a local holomorphic map  $F: (\mathbb{C}, 0) \rightarrow (\mathbb{C}, 0)$  of the form  $F(z) = \lambda z + f(z)$  with  $f(0) = f'(0) = 0$ . The question is whether there exists a local biholomorphic transformation  $\Phi: (\mathbb{C}, 0) \rightarrow (\mathbb{C}, 0)$  such that

$$\Phi \circ F = \lambda \cdot \Phi.$$

We say that  $\Phi$  linearizes  $F$  near its fixed point 0. A formal solution as in Sect. “The Normal Form Procedure” generally exists for all  $\lambda \in \mathbb{C} \setminus \{0\}$ , not equal to a root of unity. The elliptic case concerns  $\lambda$  on the complex unit circle, so of the form  $\lambda = e^{2\pi i \alpha}$ , where  $\alpha \notin \mathbb{Q}$ , and where the approximability of  $\alpha$  by rationals is of importance, cf. Siegel [76] and Bruno [34]. Siegel introduced a sufficient Diophantine condition related to (6), which by Bruno was replaced by a sufficient condition on the continued fraction approximation of  $\alpha$ . Later Yoccoz [92] proved that the latter condition is both necessary and sufficient. For a description and further comments also compare with [2,10,42,58].

*Remark*

- In certain real analytic cases Neishtadt [63], by truncating the (divergent) normal form series appropriately, obtains a remainder that is exponentially small in the perturbation parameter. Also compare with,

e. g., [2,3,4,78]. For an application in the context of the Bogdanov–Takens bifurcations for diffeomorphisms, see [9,19]. It follows that chaotic dynamics is confined to exponentially narrow horns in the parameter space.

- The growth of the Taylor coefficients of the usually divergent series of the normal form and of the normalizing transformation can be described using Gevrey symptotics [4,55,70,71,72,73]. Apart from its theoretical interest, this kind of approach is extremely useful for computational issues also compare with [4,46,78,79,80].

## Applications

We present two main areas of application of the Normal Form Theory in Perturbation Theory. The former of these deals with more globally qualitative aspects of the dynamics given by normal form approximations. The latter class of applications concerns the Averaging Theorem, where the issue is that solutions remain close to approximating solutions given by a normal form truncation.

### ‘Cantorized’ Singularity Theory

We return to the discussion of the motivation in Sect. “Motivation”, where there is a toroidal normal form symmetry up to a finite or infinite order. To begin with let us consider the quasi-periodic Hopf bifurcation [6,15], which is the analogue of the Hopf bifurcation for equilibria and the Hopf–Neimark–Sacker bifurcation for periodic solutions, in the case of quasi-periodic tori. For a description comparing the differences between all these cases we refer to [38]. For a description of the resonant dynamics in the resonant gaps see [30] and references therein. Apart from this, a lot of related work has been done in the Hamiltonian and reversible context as well, compare with [25,26,29,48].

To be more definite, we consider families of systems defined on  $\mathbb{T}^m \times \mathbb{R}^m \times \mathbb{R}^{2p} = \{x, y, z\}$ , endowed with the symplectic form  $\sigma = \sum_{j=1}^m dx_j \wedge dy_j + dz^2$ . We start with ‘integrable’, i. e.,  $x$ -independent, families of systems of the form

$$\begin{aligned}\dot{x} &= f(y, z; \lambda) \\ \dot{y} &= g(y, z; \lambda) \\ \dot{z} &= h(y, z; \lambda),\end{aligned}\tag{8}$$

compare with (5), to be considered near an invariant  $m$ -torus  $\mathbb{T}^m \times \{y_0\} \times \{z_0\}$ , meaning that we assume  $g(y_0, z_0) = 0 = h(y_0, z_0)$ . The general interest is with the persistence of such a torus under nearly integrable perturbations of (8), where we include  $\lambda$  as a multipa-

rameter. This problem belongs to the Parametrized KAM Theory [10,15,18] of which we sketch some background now. For  $y$  near  $y_0$  and  $\lambda$  near 0 consider  $\Omega(\lambda, y) = D_z h(y, z_0; \lambda)$ , noting that  $\Omega(\lambda, y) \in \text{sp}(2p, \mathbb{R})$ . Also consider the corresponding normal linear part

$$\omega(\lambda, y) \frac{\partial}{\partial x} + \Omega(\lambda, y) z \frac{\partial}{\partial z}.$$

As a first case assume that the matrix  $\Omega(0, y_0)$  has only simple non-zero eigenvalues. Then a full neighborhood of  $\Omega(0, y_0)$  in  $\text{sp}(2p, \mathbb{R})$  is completely parametrized by the eigenvalues of the matrices, where – in this symplectic case – we have to refrain from ‘counting double’. We roughly quote a KAM Theorem as this is known in the present circumstances [15]. As a nondegeneracy condition assume that the product map

$$(\lambda, y) \mapsto (\omega(\lambda, y), \Omega(\lambda, y))$$

is a submersion near  $(\lambda, y) = (0, y_0)$ . Also assume all Diophantine conditions (6), (7) to hold. Then the parametrized system (8) is *quasi-periodically stable*, which implies persistence of the corresponding Diophantine tori near  $\mathbb{T}^m \times \{y_0\} \times \{z_0\}$  under nearly integrable perturbations.

### Remark

- A key concept in the KAM Theory is that of *Whitney smoothness* of foliations of tori over nowhere dense ‘Cantor’ sets. In real analytic cases even Gevrey regularity holds, and similarly when the original setting is Gevrey; compare with [71,90]. For general reference see [10,15,18] and references therein.
- The gaps in the ‘Cantor sets’ are centered around the various resonances. Their union forms an open and dense set of small measure, where perturbation series diverge due to small divisors. In each gap the considerations mentioned at the end of Subjects. “Near a Periodic Solution” and “Near a Quasi-Periodic Torus” apply. In [25,31,33,38] the differences between period and the quasi-periodic cases are highlighted.

*Example (Quasi-periodic Hamiltonian Hopf bifurcation [26])* As a second case we take  $p = 2$ , considering the case of normal  $1 : -1$  resonance where the eigenvalues of  $\Omega(0, y_0)$  are of the form  $\pm i\mu_0$ , for a positive  $\mu_0$ . For simplicity we only consider the non-semisimple Williamson normal form

$$\Omega(0, y_0) \sim \begin{pmatrix} 0 & -\mu_0 & 1 & 0 \\ \mu_0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\mu_0 \\ 0 & 0 & \mu_0 & 0 \end{pmatrix},$$

where  $\sim$  denotes symplectic similarity. The present format of the nondegeneracy condition regarding the product map  $(\lambda, y) \mapsto (\omega(\lambda, y), \Omega(\lambda, y))$  is as before, but now it is required that the second component  $(\lambda, y) \mapsto \Omega(\lambda, y)$  is a *versal unfolding* of the matrix  $\Omega(0, y_0)$  in  $\mathfrak{sp}(4, \mathbb{R})$  with respect to the adjoint  $\mathrm{SP}(4, \mathbb{R})$ -action, compare with [2,26,27,29,44,49,54]. It turns out that a standard normalization along the lines of Sect. “[Preservation of Structure](#)” can be carried out in the  $z$ -direction, generically leading to a Hamiltonian Hopf bifurcation [88], characterized by its swallowtail geometry [86], which ‘governs’ families of invariant  $m$ -,  $(m+1)$ - and  $(m+2)$ -tori (the latter being Lagrangean). Here for the complementary spaces, see Sect. “[The Normal Form Procedure](#)”, the  $\mathfrak{sl}(2, \mathbb{R})$ -Theory is being used [39,60].

As before the question is what happens to this scenario when perturbing the system in a non-integrable way? In that case we need quasi-periodic Normal Form Theory, in the spirit of Sect. “[Semi-Local Normalization](#)”. Observe that by the  $1:-1$  resonance, difficulties occur with the third of the three Melnikov conditions (7). In a good set of Floquet coordinates the resonance can be written in the form  $\omega_j^N + \omega_\ell^N = 0$ . Nevertheless, another application of Parametrized KAM Theory [10,15,18,27] yields that the swallowtail geometry is largely preserved, when leaving out a dense union of resonance gaps of small measure. Here perturbation series diverge due to small divisors. What remains is a ‘Cantorized’ version of the swallowtail [26,27]. For a reversible analogue see [29,31].

#### Remark

- The example concerns a strong resonance and it fits in some of the larger gaps of the ‘Cantor’ set described in the former application of Parametrized KAM Theory. Apart from this, the previous remarks largely apply again. It turns out that in these and many other cases there is an infinite regression regarding the resonant bifurcation diagram.
- The combination of KAM Theory with Normal Form Theory generally has been very fruitful. In the example of Sect. “[Applications](#)” it implies that each KAM torus corresponds to a parameter value  $\lambda_0$  that is the Lebesgue density point of such quasi-periodic tori. In the real analytic case by application of [63], it can be shown that the relative measure with non-KAM tori is exponentially small in  $\lambda - \lambda_0$ . Similar results in the real analytic Hamiltonian KAM Theory (often going by the name of exponential condensation) have been obtained by [51,52,53], also com-

pare with [35,36], ► [Diagrammatic Methods in Classical Perturbation Theory](#) and with [3,10] and references therein.

#### On the Averaging Theorem

Another class of applications of normalizing-averaging is in the direction of the Averaging Theorem. There is a wealth of literature in this direction, that is not in the scope of the present paper, for further reading compare with [1,2,3,75] and references therein.

*Example (A simple averaging theorem [1])* Given is a vector field

$$\begin{aligned}\dot{x} &= \omega(y) + \varepsilon f(x, y) \\ \dot{y} &= \varepsilon g(x, y)\end{aligned}$$

with  $(x, y) \in \mathbb{T}^1 \times \mathbb{R}^n$ , compare with (4). Roughly following the recipe of the normalization process, a suitable near-identity transformation

$$(x, y) \mapsto (x, \eta)$$

of  $\mathbb{T}^1 \times \mathbb{R}^n$  yields the following reduction, after truncating at order  $O(\varepsilon^2)$ :

$$\dot{\eta} = \varepsilon \bar{g}(\eta), \quad \text{where} \quad \bar{g}(\eta) = \frac{1}{2\pi} \int_0^{2\pi} g(x, \eta) dx.$$

We now compare  $y = y(t)$  and  $\eta = \eta(t)$  with coinciding initial values  $y(0) = \eta(0)$  as  $t$  increases. The Averaging Theorem asserts that if  $\omega(\eta) > 0$  is bounded away from 0, it follows that, for a constant  $c > 0$ , one has

$$|y(t) - \eta(t)| < c\varepsilon, \quad \text{for all } t \quad \text{with } 0 \leq t \leq \frac{1}{\varepsilon}.$$

This theory extends to many classes of systems, for instance to Hamiltonian systems or, in various types of systems, in the immediate vicinity of a quasi-periodic torus. Further normalizing can produce sharper estimates that are polynomial in  $\varepsilon$ , while in the analytic case this even extends over exponentially long time intervals, usually known under the name of Nekhoroshev estimates [64,65,66], for a description and further references also see [10], ► [Nekhoroshev Theory](#). Another direction of generalization concerns passages through resonance, which in the example implies that the condition on  $\omega(\eta)$  is no longer valid. We here mention [62], for further references and descriptions referring to [2,3] and to [75].

## Future Directions

The area of research in Normal Form Theory develops in several directions, some of which are concerned with computational aspects, including the nilpotent case. Although for smaller scale projects much can be done with computer packages, for the large scale computations, e. g., needed in Celestial Mechanics, single purpose formula manipulators have to be built. For an overview of such algorithms, compare with [60,78,79,80] and references therein. Here also Gevrey asymptotics is of importance.

Another direction of development is concerned with applications in Bifurcation Theory, often in particular Singularity Theory. In many of these applications, certain coefficients in the truncated normal form are of interest and their computation is of vital importance. For an example of this in the Hamiltonian setting see [11], where the Giorgili–Galgani [45] algorithm was used to obtain certain coefficients at all arbitrary order in an efficient way. For other examples in this direction see [28,32].

Related to this is the problem how to combine the normal form algorithms as related to the present paper, with the polynomial normal forms of Singularity Theory. The latter have a universal (i. e., context independent) geometry in the product of state space and parameter space. A problem of relevance for applications is to pull-back the Singularity Theory normal form back to the original system. For early attempts in this direction see [20,23], which, among other things, involve Gröbner basis techniques.

## Bibliography

### Primary Literature

- Arnold VI (1980) *Mathematical methods of classical mechanics*. Springer, New York (English, Russian original)
- Arnold VI (1983) *Geometrical methods in the theory of ordinary differential equations*. Springer, New York (English, Russian original)
- Arnold VI (ed) (1988) *Dynamical systems III, encyclopædia of mathematical sciences, vol 3*. Springer, Berlin (English, Russian original)
- Baldomá I, Haro A (2008) One dimensional invariant manifolds of Gevrey type in real-analytic maps. *DCDS series B* 10(2&3):295–322
- Birkhoff GD (1927) *Dynamical systems*. AMS Publications
- Braaksma BLJ, Broer HW (1987) On a quasi-periodic Hopf bifurcation. *Ann Inst Henri Poincaré, Analyse Non Linéaire* 4(2):115–168
- Broer HW (1981) Formal normal form theorems for vector fields and some consequences for bifurcations in the volume preserving case. In: Rand DA, Young LS (eds) *Dynamical systems and turbulence*. Warwick, 1980 LNM 898. Springer, Berlin, pp 54–74
- Broer HW (1993) Notes on perturbation theory 1991. In: Erasmus ICP (ed) *Mathematics and Fundamental Applications*. Aristotle University Thessaloniki, pp 44
- Broer HW, Roussarie R (2001) Exponential confinement of chaos in the bifurcation set of real analytic diffeomorphisms. In: Broer HW, Krauskopf B, Vegter G (eds) *Global analysis of dynamical systems. Festschrift dedicated to Floris Takens for his 60th birthday* Bristol Phila IOP, pp 167–210
- Broer HW, Sevryuk MB (2008) *kam Theory: quasi-periodicity in dynamical systems*. In: Broer HW, Hasselblatt B, Takens F (eds) *Handbook of Dynamical Systems, vol 3*. North-Holland (to appear 2009)
- Broer HW, Simó C (2000) Resonance tongues in Hill's equations: a geometric approach. *J Differ Equ* 166:290–327
- Broer HW, Takens F (1989) Formally symmetric normal forms and genericity. *Dyn Rep* 2:36–60
- Broer HW, Tangerman FM (1986) From a differentiable to a real analytic perturbation theory, applications to the Kupka Smale theorems. *Ergod Theor Dyn Syst* 6:345–362
- Broer HW, Vegter G (1992) Bifurcational aspects of parametric resonance. *Dyn Rep New Ser* 1:1–51
- Broer HW, Huitema GB, Takens F, Braaksma BLJ (1990) Unfoldings and bifurcations of quasi-periodic tori. *Mem AMS* 83(421):i–vii; pp 1–175
- Broer HW, Dumortier F, van Strien SJ, Takens F (1991) Structures in dynamics: Finite dimensional deterministic studies. In: van Groesen EWC, de Jager EM (eds) *Studies in mathematical physics, vol 2*. North-Holland, Amsterdam (Russian translation 2003)
- Broer HW, Chow SN, Kim Y, Vegter G (1993) A normally elliptic Hamiltonian bifurcation. *ZAMP* 44:389–432
- Broer HW, Huitema GB, Sevryuk MB (1996) Quasi-periodic motions in families of dynamical systems: Order amidst chaos. *LNM 1645* Springer, Berlin
- Broer HW, Roussarie R, Simó C (1996) Invariant circles in the Bogdanov-Takens bifurcation for diffeomorphisms. *Ergod Theor Dyn Syst* 16:1147–1172
- Broer HW, Hoveijn I, Lunter GA, Vegter G (1998) Resonances in a Spring–Pendulum: algorithms for equivariant singularity theory. *Nonlinearity* 11(5):1–37
- Broer HW, Simó C, Tatjer JC (1998) Towards global models near homoclinic tangencies of dissipative diffeomorphisms. *Nonlinearity* 11:667–770
- Broer HW, Takens F, Wagener FOO (1999) Integrable and non-integrable deformations of the skew Hopf bifurcation. *Reg Chaotic Dyn* 4(2):17–43
- Broer HW, Hoveijn I, Lunter GA, Vegter G (2003) Bifurcations in Hamiltonian systems: Computing singularities by Gröbner bases. *LNM, vol 1806*. Springer, Berlin
- Broer HW, Golubitsky M, Vegter G (2003) The geometry of resonance tongues: A singularity theory approach. *Nonlinearity* 16:1511–1538
- Broer HW, Hanßmann H, Jorba A, Villanueva J, Wagener FOO (2003) Normal-internal resonances in quasi-periodically forced oscillators: a conservative approach. *Nonlinearity* 16:1751–1791
- Broer HW, Hanßmann H, Hoo J (2007) The quasi-periodic Hamiltonian Hopf bifurcations. *Nonlinearity* 20:417–460
- Broer HW, Hoo J, Naudot V (2007) Normal linear stability of quasi-periodic tori. *J Differ Equ* 232(2):355–418



28. Broer HW, Golubitsky M, Vegter G (2007) Geometry of resonance tongues. In: Chéniot D, Dutertre N, Murolo C, Trotman D, Pichon A (eds) *Proceedings of the 2005 marseille singularity school and conference, dedicated to Jean-Paul Brasselet on his 60th birthday*. World Scient, pp 327–356
29. Broer HW, Ciocci MC, Hanßmann H (2007) The quasi-periodic reversible Hopf bifurcation. *IJBC* 17(8):2605–2623
30. Broer HW, Naudot V, Roussarie R, Saleh K, Wagener FOO (2007) Organising centres in the semi-global analysis of dynamical systems. *IJAMAS* 12(D07):7–36
31. Broer HW, Ciocci MC, Hanßmann H, Vanderbauwhede A (2008) Quasi-periodically stable unfoldings of normally resonant tori. *Physica D* (to appear)
32. Broer HW, Vegter G (2008) Generic Hopf-Neimark-Sacker bifurcations in feed forward systems. *Nonlinearity* 21:1547–1578
33. Broer HW, Hanßmann H, You J ( ) On the destruction of resonant Lagrangean tori in Hamiltonian systems. (in preparation)
34. Bruno AD (1965) On the convergence of transformations of differential equations to the normal form. *Soviet Math Dokl* 6:1536–1538 (English, Russian original)
35. Bruno AD (1989) *Local methods in nonlinear differential equations*. Springer, Berlin (English, Russian original)
36. Bruno AD (2000) *Power geometry in algebraic and differential equations*. North-Holland Mathematical Library, vol 57. North-Holland, Amsterdam (English, Russian original)
37. Chow SN, Li C, Wang D (1994) *Normal forms and bifurcations of planar vector fields*. Cambridge University Press, Cambridge
38. Ciocci MC, Litvak-Hinenzon A, Broer HW (2005) Survey on dissipative KAM theory including quasi-periodic bifurcation theory based on lectures by Henk Broer. In: Montaldi J, Ratiu T (eds) *Geometric mechanics and symmetry: The Peyresq lectures*. LMS Lecture Notes Series, vol 306. Cambridge University Press, Cambridge, pp 303–355
39. Cushman RH, Sanders JA (1986) Nilpotent normal forms and representation theory of  $sl(2, \mathbb{R})$ . In: Golubitsky M, Guckenheimer J (eds) *Multi-parameter bifurcation theory* AMS. Providence, pp 31–51
40. Deprit A (1969) Canonical transformations depending on a small parameter. *Celest Mech* 1:12–30
41. Deprit A (1981) The elimination of the parallax in satellite theory. *Celest Mech* 24:111–153
42. Eliasson LH, Kuksin SB, Marmi S, Yoccoz JC (2002) *Dynamical systems and small divisors*. LNM, vol 1784. Springer, Berlin
43. Ferrer S, Hanßmann H, Palacián J, Yanguas P (2002) On perturbed oscillators in 1-1-1 resonance: the case of axially symmetric cubic potentials. *J Geom Phys* 40:320–369
44. Galin DM (1982) Versal deformations of linear Hamiltonian systems. *Am Soc Trans Ser 2*(118):1–12
45. Giorgilli A, Galgani L (1978) Formal integrals for an autonomous Hamiltonian system near an equilibrium point. *Celest Mech* 17:267–280
46. 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 Differ Equ* 77:167–198
47. Gustavson FG (1966) On constructing formal integrals of a Hamiltonian system near an equilibrium point. *Astron J* 71:670–686
48. Hanßmann H (2007) Local and semi-local bifurcations in hamiltonian dynamical systems. *LNM*, vol 1893. Springer, Berlin
49. Hoveijn I (1996) Versal deformations and normal forms for reversible and Hamiltonian linear systems. *J Diff Eqns* 126(2):408–442
50. Ioss G (1988) Global characterization of the normal form of a vector field near a closed orbit. *J Diff Eqns* 76:47–76
51. Jorba À, Villanueva J (1997) On the persistence of lower dimensional invariant tori under quasi-periodic perturbations. *J Nonlinear Sci* 7:427–473
52. Jorba À, Villanueva J (1997) On the normal behaviour of partially elliptic lower-dimensional tori of Hamiltonian systems. *Nonlinearity* 10:783–822
53. Jorba À, Villanueva J (2001) The fine geometry of the Cantor families of invariant tori in Hamiltonian systems. In: Casacuberta C, Miró-Roig RM, Verdera J, Xambó-Descamps S (eds) *European Congress of Mathematics, Barcelona, 2000*, vol 2. *Progress in Mathematics*, vol 202. Birkhäuser, Basel, pp 557–564
54. Koçak H (1984) Normal forms and versal deformation of linear Hamiltonian systems. *J Diff Eqns* 51:359–407
55. Marco JP, Sauzin D (2002) Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian systems. *Publ IHES* 96:199–275
56. Marsden JE, McCracken M (1976) *The Hopf-bifurcation and its applications*. Springer, New York
57. Meyer KR (1984) Normal forms for the general equilibrium. *Funkcial Ekv* 27:261–271
58. Milnor JW (2006) *Dynamics in one complex variable*, 3rd edn. *Ann Math Stud*, vol 160. Princeton Univ Press, Princeton
59. Moser JK (1968) Lectures on Hamiltonian systems. *Mem Amer Math Soc* 81:1–60
60. Murdock J (2003) *Normal forms and unfoldings for local dynamical systems*. Springer Monographs in Mathematics. Springer, New York
61. Narasimhan R (1968) *Analysis on real and complex manifolds*. Elsevier, North Holland
62. Neishtadt AI (1975) Passage through resonances in two-frequency problem. *Sov Phys Dokl* 20(3):189–191 (English, Russian original)
63. Neishtadt AI (1984) The separation of motions in systems with rapidly rotating phase. *J Appl Math Mech* 48(2):133–139 (English, Russian original)
64. Nekhoroshev NN (1971) On the behavior of Hamiltonian systems close to integrable ones. *Funct Anal Appl* 5:338–339 (English, Russian original)
65. Nekhoroshev NN (1977) An exponential estimate of the stability time of nearly integrable Hamiltonian systems. I, Russian Math Surv 32(6):1–65 (English, Russian original)
66. Nekhoroshev NN (1979) An exponential estimate of the stability time of nearly integrable Hamiltonian systems, vol II. *Trudy Sem Imeni*. In: Petrovskogo IG pp 5:5–50 (in Russian) English translation. In: Oleinik OA (ed) (1985) *Topics in modern mathematics*, Petrovskii Seminar. Consultants Bureau, New York, pp 5:1–58
67. Palis J de Melo WC (1982) *Geometric Theory of Dynamical Systems*. Springer
68. Palis J, Takens F (1993) *Hyperbolicity & sensitive chaotic dynamics at homoclinic bifurcations*. Cambridge Studies in Advanced Mathematics, Cambridge University Press 35
69. Poincaré H (1928) *Œuvres*, vol I. Gauthier-Villars
70. Popov G (2000) Invariant tori, effective stability, and quasi-modes with exponentially small error terms. I: Birkhoff normal forms. *Ann Henri Poincaré* 1(2):223–248

71. Popov G (2004) KAM theorem for Gevrey Hamiltonians. *Ergod Theor Dynam Syst* 24:1753–1786
72. Ramis JP (1994) Séries divergentes et théories asymptotiques. *Panor Synth* pp 0–74
73. Ramis JP, Schäfke R (1996) Gevrey separation of slow and fast variables. *Nonlinearity* 9:353–384
74. Roussarie R (1987) Weak and continuous equivalences for families of line diffeomorphisms. In: *Dynamical systems and bifurcation theory*, Pitman research notes in math. Series Longman 160:377–385
75. Sanders JA, Verhulst F, Murdock J (1985) *Averaging methods in nonlinear dynamical systems*. Revised 2nd edn, Appl Math Sciences 59, 2007. Springer
76. Siegel CL (1942) Iteration of analytic functions. *Ann Math* 43(2):607–612
77. Siegel CL, Moser JK (1971) *Lectures on celestial mechanics*. Springer, Berlin
78. Simó C (1994) Averaging under fast quasiperiodic forcing. In: Seimenis J (ed) *Hamiltonian mechanics, integrability and chaotic behaviour*. NATO Adv Sci Inst Ser B Phys 331:13–34, Plenum, New York
79. Simó C (1998) Effective computations in celestial mechanics and astrodynamics. In: Rumyantsev VV, Karapetyan AV (eds) *Modern methods of analytical mechanics and their applications*. CISM Courses Lectures, vol 387. Springer, pp 55–102
80. Simó C (2000) Analytic and numeric computations of exponentially small phenomena. In: Fiedler B, Gröger K, Sprekels J (eds) *Proceedings EQUADIFF 99*, Berlin. World Scientific, Singapore, pp 967–976
81. Spivak M (1970) *Differential Geometry*, vol I. Publish or Perish Inc
82. Sternberg S (1959) On the structure of local homeomorphisms of Euclidean  $n$ -space, vol II. *Amer J Math* 81:578–605
83. Takens F (1973) Forced oscillations and bifurcations. *Applications of global analysis*, vol I, Utrecht. Comm Math Inst Univ Utrecht, pp 31–59 (1974). Reprinted In: Broer HW, Krauskopf B, Vegter G (eds) (2001) *Global analysis of dynamical systems*. Festschrift dedicated to Floris Takens for his 60th birthday Leiden. Inst Phys Bristol, pp 1–61
84. Takens F (1974) Singularities of vector fields. *Publ Math IHÉS* 43:47–100
85. Takens F, Vanderbauwhede A (2009) Local invariant manifolds and normal forms. In: Broer HW, Hasselblatt B, Takens F (eds) *Handbook of dynamical systems*, vol 3. North-Holland (to appear)
86. Thom R (1989) *Structural stability and morphogenesis*. An outline of a general theory of models, 2nd edn. Addison-Wesley, Redwood City (English, French original)
87. Vanderbauwhede A (1989) Centre manifolds, normal forms and elementary bifurcations. *Dyn Rep* 2:89–170
88. van der Meer JC (1985) The hamiltonian Hopf bifurcation. *LNM*, vol 1160. Springer
89. Varadarajan VS (1974) Lie groups, Lie algebras and their representations. Englewood Cliffs, Prentice-Hall
90. Wagener FOO (2003) A note on Gevrey regular KAM theory and the inverse approximation lemma. *Dyn Syst* 18(2):159–163
91. Wagener FOO (2005) On the quasi-periodic  $d$ -fold degenerate bifurcation. *J Differ Eqn* 216:261–281
92. Yoccoz JC (1995) Théorème de Siegel, nombres de Bruno et polynômes quadratiques. *Astérisque* 231:3–88

## Books and Reviews

- Braaksma BLJ, Stolovitch L (2007) Small divisors and large multipliers (Petits diviseurs et grands multiplicateurs). *Ann l'institut Fourier* 57(2):603–628
- Broer HW, Levi M (1995) Geometrical aspects of stability theory for Hill's equations. *Arch Rat Mech An* 131:225–240
- Gaeta G (1999) Poincaré renormalized forms. *Ann Inst Henri Poincaré* 70(6):461–514
- Martinet J, Ramis JP (1982) Problèmes des modules pour les équations différentielles non linéaires du premier ordre. *Publ IHES* 55:63–164
- Martinet J, Ramis JP (1983) Classification analytique des équations différentielles non linéaires résonnantes du premier ordre. *Ann Sci École Norm Supérieure Sér 4* 16(4):571–621
- Vanderbauwhede A (2000) Subharmonic bifurcation at multiple resonances. In: Elaydi S, Allen F, Elkhader A, Mughrabi T, Saleh M (eds) *Proceedings of the mathematics conference*, Birzeit, August 1998. World Scientific, Singapore, pp 254–276

## Numerical Bifurcation Analysis

HIL MEIJER<sup>1</sup>, FABIO DERCOLE<sup>2</sup>, BART OLDEMAN<sup>3</sup>

<sup>1</sup> Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands

<sup>2</sup> Department of Electronics and Information, Politecnico di Milano, Milano, Italy

<sup>3</sup> Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Continuation and Discretization of Solutions](#)

[Normal Forms and the Center Manifold](#)

[Continuation and Detection of Bifurcations](#)

[Branch Switching](#)

[Connecting Orbits](#)

[Software Environments](#)

[Future Directions](#)

[Bibliography](#)

## Glossary

**Dynamical system** A rule for time evolution on a state space. The term system will be used interchangeably. Here a system is a family given by an ordinary differential equation (ODE) depending on parameters.

**Equilibrium** A constant solution of the system, for given parameter values.

**Limit cycle** An isolated periodic solution of the system, for given parameter values.

**Bifurcation** A qualitative change in the dynamics of a dynamical system produced by changing its parameters. Bifurcation points are the critical parameter combinations at which this happens for arbitrarily small parameter perturbations.

**Normal form** A simplified model system for the analysis of a certain type of bifurcation.

**Codimension** The minimal number of parameters needed to perturb a family of systems in a generic manner.

**Defining system** A set of suitable equations so that the zero set corresponds to a bifurcation of a certain type or to a particular solution of the system. Also called defining function or equation.

**Continuation** A numerical method suited for tracing one-dimensional manifolds, curves (here called branches) of solutions for a defining system while one or more parameters are varied.

**Test function** A function designed to have a regular zero at a bifurcation. During continuation a test function can be monitored to detect bifurcations.

**Branch switching** Several branches of different codimension can emanate from a bifurcation point. Switching from the computation of one branch to an other requires appropriate procedures.

### Definition of the Subject

The theory of dynamical systems studies the behavior of solutions of systems, like nonlinear ordinary differential equations (ODEs), depending upon parameters. Using qualitative methods of bifurcation theory, the behavior of the system is characterized for various parameter combinations. In particular, the catalog of system behaviors showing qualitative differences can be identified, together with the regions in parameter space where the different behaviors occur. Bifurcations delimit such regions. Symbolic and analytical approaches are in general infeasible, but numerical bifurcation analysis is a powerful tool that aids in the understanding of a nonlinear system. When computing power became widely available, algorithms for this type of analysis matured and the first codes were developed. With the development of suitable algorithms, the advancement in the qualitative theory has found its way into several software projects evolving over time. The availability of software packages allows scientists to study and adjust their models and to draw conclusions about their dynamics.

### Introduction

Nonlinear ordinary differential equations depending upon parameters are ubiquitous in science. In this article meth-

ods for *numerical bifurcation analysis* are reviewed, an approach to investigate the dynamic behavior of nonlinear dynamical systems given by

$$\dot{x} = f(x, p), \quad x \in \mathbb{R}^n, \quad p \in \mathbb{R}^{n_p}, \quad (1)$$

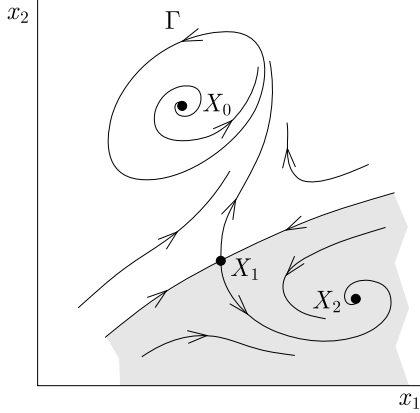
where  $f: \mathbb{R}^n \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}^n$  is generic and sufficiently smooth. In particular,  $x(t)$  represents the state of the system at time  $t$  and its components are called *state* (or *phase variables*),  $\dot{x}(t)$  denotes the time derivative of  $x(t)$ , while  $p$  denotes the *parameters* of the system, representing experimental control settings or variable inputs.

In many instances, solutions of (1), starting at an initial condition  $x(0)$ , appear to converge as  $t \rightarrow \infty$  to *equilibria* (steady states) or *limit cycles* (periodic orbits). Bounded solutions can also converge to more complex *attractors*, like *tori* (quasi-periodic orbits) or *strange attractors* (chaotic orbits). The attractors of the system are invariant under time evolution, i. e., under the application of the time- $t$  map  $\Phi^t$ , where  $\Phi$  denotes the flow induced by the system (1). Solutions attracting all nearby initial conditions, are said to be *stable*, and *unstable* if they repel some initial conditions.

Generally speaking, it is hard to obtain closed formulas for  $\Phi^t$  as the system is nonlinear. In some cases, one can compute equilibria analytically, but this is often not the case for limit cycles. However, numerical simulations of (1) easily give an idea of how solutions look, although one never computes the true orbit due to numerical errors. One can verify stability conditions by linearizing the flow around equilibria and cycles. In particular, an equilibrium  $x_0$  is stable if the eigenvalues of the linearization (*Jacobi*) matrix  $A = f_x(x_0, p)$  (where the subscript denotes differentiation) all have a negative real part. Similarly, for a limit cycle  $x_0(t)$  with period  $T$ , one defines the *Floquet multipliers* (or simply multipliers) as the eigenvalues of the monodromy matrix  $M = \Phi_x^T(x_0(0))$ . The cycle is stable if the nontrivial multipliers, there is always one equal to 1, are all within the unit circle. Equilibria and limit cycles are called *hyperbolic* if the eigenvalues and nontrivial multipliers do not have a zero real part or modulus one, respectively.

For any given parameter combination  $p$ , the state space representation of all orbits constitutes the *phase portrait* of the system. In practice, one draws a set of strategic orbits (or finite segments of them), from which all other orbits can be intuitively inferred, as illustrated in Fig. 1.

Points  $X_0, X_1, X_2$  are equilibria, of which  $X_0$  and  $X_1$  are unstable and  $X_2$  is stable. In particular,  $X_0$  is a *repellor*, i. e., nearby orbits do not tend to remain close to  $X_0$ , while  $X_1$  is a *saddle*, i. e., almost all nearby orbits go away from  $X_1$  except two, which tend to  $X_1$  and lie on the so-called *stable manifold*; the two orbits emanating from  $X_1$



**Numerical Bifurcation Analysis, Figure 1**

Phase portrait of a two-dimensional system with two attractors (the equilibrium  $X_2$  and the limit cycle  $\Gamma$ ), a repeller ( $X_0$ ), and a saddle ( $X_1$ )

compose the *unstable manifold*. There are therefore two attractors, the equilibrium  $X_2$  and the limit cycle  $\Gamma$ , whose basins of attraction consist of the initial conditions in the shaded and white areas, respectively. Note that while attractors and repellers can be easily obtained through simulation, forward and backward in time, saddles can be hard to find.

The analysis of system (1) becomes even more difficult if one wants to follow the phase portrait under variation of parameters. Generically, by perturbing a parameter slightly the phase portrait changes slightly as well. Namely, if the new phase portrait is topologically equivalent to the original one, then nothing changed from a qualitative point of view, i. e., all attracting, repelling, and saddle sets are still present with unchanged stability properties, though slightly perturbed. By contrast, the critical points in parameter space where arbitrarily small parameter perturbations give rise to nonequivalent phase portraits are called *bifurcation points*, where bifurcations are said to occur. Bifurcations therefore result in a partition of parameter space into regions: parameter combinations in the same region correspond to topologically equivalent dynamics, while nonequivalent phase portraits arise for parameter combinations in neighboring regions. Most often, this partition is represented by means of a two-dimensional *bifurcation diagram*, where the regions of a parameter plane are separated by so-called *bifurcation curves*. Bifurcations are said to be *local* if they occur in an arbitrarily small neighborhood of the equilibrium or cycle; otherwise, they are said to be *global*.

Although one might hope to detect bifurcations by simulating system (1) for various parameter combinations and initial conditions, a “brute force” simulation approach

is hardly effective and accurate in practice, because bifurcations of equilibria and cycles are associated with a loss of hyperbolicity, e. g., stability, so that one should dramatically increase the length of simulations while approaching the bifurcation. In particular, saddle sets are hard to find by simulation, but play a fundamental role in bifurcation analysis, since they, together with attracting and repelling sets, form the skeleton of the phase portrait. This is why numerical bifurcation analysis does not rely on simulation, but rather on *continuation*, a numerical method suited for computing (approximating through a discrete sequence of points) one-dimensional manifolds (curves, “branches” in regular) implicitly defined as the zero set of a suitable *defining function*.

The general idea is to formulate the computation of equilibria and their bifurcations as a suitable *algebraic problem* (AP) of the form

$$F(u, p) = 0, \quad (2)$$

where  $u \in \mathbb{R}^{n_u}$  is composed of  $x$  and possibly other variables characterizing the system, see, e. g., defining functions as in Sect. “Continuation and Detection of Bifurcations”. Here, however, for simplicity of notation,  $u$  will be considered as in  $\mathbb{R}^n$ , but the actual dimension of  $u$  will always be clear from the context. Similarly, limit cycles and their bifurcations are formulated in the form of a *boundary-value problem* (BVP)

$$\begin{cases} \dot{u} - f(u, p) = 0, \\ g(u(0), u(T), p) = 0, \\ \int_0^T h(u(t), p) dt = 0, \end{cases} \quad (3)$$

with  $n_b$  boundary conditions, i. e.,  $g: \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{n_b}$ ,  $n_i$  integral conditions, i. e.,  $h: \mathbb{R}^n \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{n_i}$ , and  $u$  in a proper function space. In other words, a list of defining functions is formulated, in the form (2) or (3), to cover all cases of interest. For example,  $u = x$  and  $F(x, p) = f(x, p)$  is the AP defining equilibria of (1). The commonly used cycle BVP, with the time-rescaling  $t = T\tau$ , is

$$\begin{cases} \dot{x} - Tf(x, p) = 0, \\ x(0) - x(1) = 0, \\ \int_0^1 x(\tau)^\top \dot{x}^{k-1}(\tau) d\tau = 0, \end{cases} \quad (4)$$

where from here on  $^\top$  denotes the transpose and, for simplicity,  $\cdot$  for  $d/d\tau$  is used. The integral condition is the so-called *phase condition*, which ensures that  $x(\tau)$  is the 1 periodic solution of (1) closest to the reference solution  $x^{k-1}(\tau)$  (typically known from the previous point along the continuation), among time-shifted orbits  $x(\tau - \tau_0)$ ,  $\tau_0 \in [0, 1]$ .



As will be discussed in Sect. “Discretization of BVPs”, a proper time-discretization of  $u(\tau)$  allows one to approximate any BVP by a suitable AP. Thus, equilibria, limit cycles and their bifurcations can all be represented by an algebraic defining function like (2), and numerical continuation allows one to produce one-dimensional solution branches of (2) under the variation of strategic components of  $p$ , called *free* parameters. With this approach, equilibria and cycles can be followed without further difficulty in parameter regimes where these are unstable. Then, during continuation, the stability of equilibria and cycles is determined through linearization. Moreover, the characterization of nearby solutions of (1) can be done using *normal forms*, i. e., the simplest canonical models to which the system, close to a bifurcation, can be reduced on a lower-dimensional manifold of the state space, the so-called *center manifold*. While a branch of equilibria or cycles is followed, bifurcations can be detected as the zero of suitable test functions. Upon detection of a bifurcation, the defining function can be augmented by this test function or another appropriate function, and the new defining function can then be continued using one more free parameter.

An analytical bifurcation study is feasible for simple systems only. Numerical bifurcation analysis is one of the few but also very powerful tools to understand and describe the dynamics of systems depending on parameters. Some basic steps while performing bifurcation analysis will be outlined and software implementations of continuation and bifurcation algorithms discussed.

First a few standard and often used approaches for the computation and continuation of zeros of a defining function are reviewed in Sect. “Continuation and Discretization of Solutions”. The presentation starts with the most obvious, but also naive, approaches to contrast these with the methods employed by software packages. In Sect. “Normal Forms and the Center Manifold” several possible scenarios for the loss of stability of equilibria and limit cycles are discussed. Not all bifurcations are characterized by linearization and for the detection and analysis of these bifurcations, codimension 1 normal forms are mentioned and a general method for their computation on a center manifold is presented. Then, a list of suitable test functions and defining systems for the computation of bifurcation branches is discussed in Sect. “Continuation and Detection of Bifurcations”. In particular, when a system bifurcates new solution branches appear. Techniques to switch to such new branches are described in Sect. “Branch Switching”. Finally, the computation and continuation of global bifurcations characterized by orbits connecting equilibria is presented, in particular *homoclinic* orbits, in Sect. “Connecting Orbits”. This review concludes

with an overview of existing implementations of the described algorithms in Sect. “Software Environments”. Previous reviews [7,9,22,43] have similar contents. This review however, focuses more on the principles now underlying the most frequently used software packages for bifurcation analysis and the algorithms being used.

## Continuation and Discretization of Solutions

The continuation of a solution  $u$  of (2) with respect to one parameter  $p$  is a fundamental application of the Implicit Function Theorem (IFT).

Generally speaking, to define one-dimensional solution manifolds (branches), the number of unknowns in (2) should be one more than the number of equations, i. e.,  $n_p = 1$ . However, during continuation it is better not to distinguish between state variables and parameters as will become apparent in Sects. “Pseudo-Arclength Continuation” and “Moore–Penrose Continuation”. Therefore, write  $y = (u, p) \in Y = \mathbb{R}^{n+1}$  for the continuation variables in the continuation space  $Y$  and consider the continuation problem

$$F(y) = 0, \quad (5)$$

with  $F: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ .

Let  $F$  be at least continuous differentiable,  $y_0 = (u_0, p_0)$  be a known solution point of (5), and the matrix  $F_y(y_0) = [F_u(u_0, p_0) | F_p(u_0, p_0)]$  be full rank, i. e.,

$$\text{rank}(F_y(y_0)) = n \Leftrightarrow \begin{cases} \text{(i) } \text{rank}(F_u(u_0, p_0)) = n, \text{ or} \\ \text{(ii) } \text{rank}(F_u(u_0, p_0)) = n-1 \text{ and} \\ \quad F_p(u_0, p_0) \notin \mathcal{R}(F_u(u_0, p_0)), \end{cases} \quad (6)$$

where  $\mathcal{R}(F_u)$  denotes the range of  $F_u$ . Then the IFT states that there exists a unique solution branch of (5) locally to  $y_0$ . Introducing a scalar coordinate  $s$  parametrizing the branch, e. g., the *arclength* positively measured from  $y_0$  in one of the two directions along the solution branch, then one can represent the branch by  $y(s) = (u(s), p(s))$  and the IFT guarantees that  $F(y(s)) = 0$  for  $|s|$  is sufficiently small. Moreover,  $y(s)$  is continuous differentiable and the vector  $\phi(s) = y_s(s) = (u_s(s), p_s(s)) = (v(s), q(s))$ , tangent to the solution branch at  $y(s)$ , exists and is the unique solution of

$$\begin{cases} F_y(y(s))\phi(s) = F_u(u(s), p(s))v(s) + F_p(u(s), p(s))q(s) = 0, \\ \phi(s)^\top \phi(s) = v(s)^\top v(s) + q(s)^2 = 1. \end{cases} \quad (7)$$



In other words, the matrix  $F_y(y_0)$  has a one-dimensional nullspace  $\mathcal{N}(F_y(y_0))$  spanned by  $\phi(0)$  and  $y_0$  is said to be a *regular* point of the continuation space  $Y$ .

Below, several variants of numerical continuation will be described. The aim is to produce a sequence of points  $y_k$ ,  $k \geq 0$  that approximate the solution branch  $y(s)$  in one direction. Starting from  $y_0$ , the general idea is to make a suitable prediction  $y_1^0$ , typically along the tangent vector, from which the Newton method is applied to find the new point  $y_1$ . The predictor-corrector procedure is then iterated. First, the simplest implementation is presented and it is shown where it might fail. Many continuation packages for bifurcation theory use an alternative implementation of which two variants are discussed. Many more advanced predictor-corrector schemes have been designed, see [1,18,46] and references therein.

### Parameter Continuation

Parameter continuation assumes that the solution branch of (5) can be parameterized by the parameter  $p \in \mathbb{R}$ . Indeed, if  $F_u$  has full rank, i.e., case (i) in (6), then this is possible by the IFT. Starting from  $(u_0, p_0)$  and perturbing the parameter a little, with a *stepsize*  $h$ , the new parameter is  $p_1 = p_0 + h$  and the most simple predictor for the state variable is given by  $u_1^0 = u_0$ .

Application of Newton's method to find  $u_1$  satisfying (5) leads to

$$u_1^{j+1} = u_1^j - F_u(u_1^j, p_1)^{-1} F(u_1^j, p_1), \quad j = 0, 1, 2, \dots$$

The iterations are stopped when a certain accuracy is achieved, i.e.,  $\|\Delta u\| = \|u_1^{j+1} - u_1^j\| < \varepsilon_u$  and/or  $\|F(u_1^j, p_1)\| < \varepsilon_F$ . In practice, also the maximum number

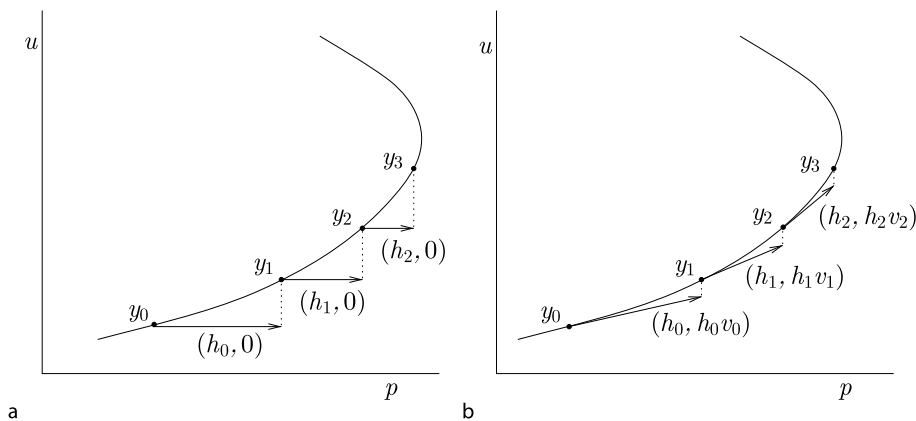
of Newton steps is bounded, in order to guarantee termination. If this maximum is reached before convergence, the computation is restarted with a smaller (typically halved) stepsize. However, in case of quick convergence, e.g., after only a few iterations, the stepsize is multiplied (1.3 is a typical factor). In any case, the stepsize is varied between two assigned limits  $h_{\min}$  and  $h_{\max}$ , so that continuation cannot proceed when convergence is not reached even with minimum stepsize. When  $h$  is chosen too small, too much computational work may be performed, while for  $h$  that is chosen too large, little detail of the solution branch is obtained.

As a first improved predictor, note that the IFT suggests to use the tangent prediction for the state variables  $u_1^0 = u_0 + hv_0$ , where  $v_0$  is obtained from (7) with  $s = 0$  and as  $v(0)/q(0)$ . Indeed, the tangent vector can be approximated by the difference  $v_k = (u_k - u_{k-1})/h_k$  or, even better, computed at negligible cost, since the numerical decomposition of the matrix  $F_u(u_k, p_k)$  is known from the last Newton iteration.

These methods are illustrated in Fig. 2. Note in particular the folds in the sketch. Here parameter continuation does not work, since exactly at the fold  $F_u(u, p)$  is singular such that Newton's method does not converge and beyond, for larger  $p$ , there is no local solution to (5).

### Pseudo-Arclength Continuation

Near folds, the solution branch is not well parameterized by the parameter, but one can use a state variable for the parametrization. In fact, the fold is a regular point (case (ii) in (6)) at which the tangent vector  $\phi = (v, q)$  has no parameter component, i.e.,  $q = 0$ . So, without distinguishing between parameters and state variables, one takes the



Numerical Bifurcation Analysis, Figure 2

Parameter continuation without (a) and with (b) tangent prediction. The *dotted lines* indicate subspaces where solutions are searched

tangent prediction  $y_1^0 = y_0 + h\phi_0$ , as long as the starting solution  $y_0$  is a regular point. Since now both  $p$  and  $u$  are corrected, one more constraint is needed. Pseudo-arclength continuation uses the stepsize  $h$  as an approximation of the required distance, in arclength, between  $y_0$  and the next point  $y_1$ . This leads to the so-called *pseudo-arclength equation*  $\phi_0^\top(y_1 - y_0) = h$ . In this way, solution branches can be followed past folds. The idea for this continuation method is due to Keller [50].

The Newton iteration, applied to

$$\begin{cases} F(y_1) = 0, \\ \phi_0^\top(y_1 - y_0) - h = 0, \end{cases}$$

is given by

$$y_1^{j+1} = y_1^j - \left( cF_y(y_1^j) \right)^{-1} \begin{pmatrix} cF(y_1^j) \\ 0 \end{pmatrix}, \quad j = 0, 1, 2, \dots, \quad (8)$$

where  $\Delta y = y_1^{j+1} - y_1^j$  is forced to lie in the hyperplane orthogonal to the tangent vector, as illustrated in Fig. 3a. Upon convergence, the new tangent vector  $\phi_1$  is obtained by solving (7) at  $y_1$ .

### Moore–Penrose Continuation

This continuation method is based on optimization. Starting with the tangent prediction  $y_1^0 = y_0 + h\phi_0$ , a point  $y_1$  with  $F(y_1) = 0$  nearest to  $y_1^0$  is searched, so the following is optimized

$$\min_{y_1} \{ \|y_1 - y_1^0\| \mid F(y_1) = 0 \}.$$

Each correction is therefore required to be orthogonal to the nullspace of  $F_y(y_1^j)$ , i. e.,

$$\begin{cases} F(y_1^{j+1}) = 0, \\ (\phi_1^j)^\top(y_1^{j+1} - y_1^j) = 0. \end{cases}$$

Starting with  $\phi_1^0 = \phi_0$  the Newton iterations are given by

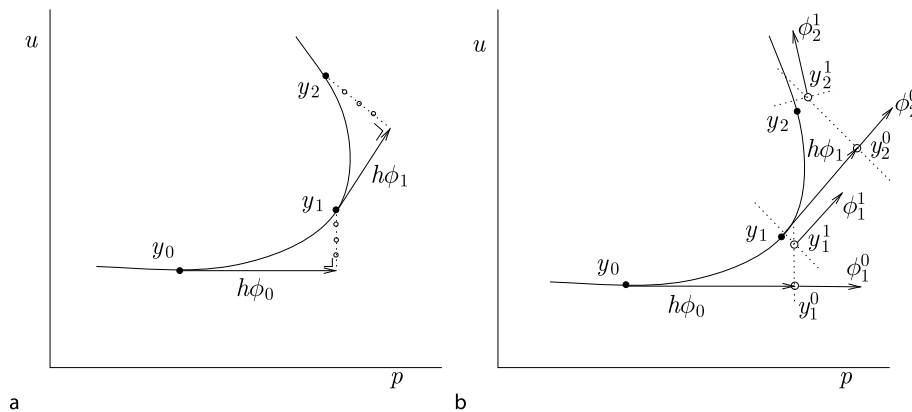
$$\begin{cases} y_1^{j+1} = y_1^j - \left( F_y(y_1^j) \right)^{-1} \begin{pmatrix} F(y_1^j) \\ 0 \end{pmatrix}, \\ \phi_1^{j+1} = \left( F_y(y_1^{j+1}) \right)^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad j = 0, 1, 2, \dots \end{cases} \quad (9)$$

As illustrated in Fig. 3b, the Moore–Penrose continuation can be interpreted as a variant of Keller’s method in which the tangent vector is updated at every Newton step. When the new point  $y_1$  is found, the tangent vector  $\phi_1$  is immediately obtained as  $\phi_1^{j+1}/\|\phi_1^{j+1}\|$  from the last Newton iteration, since  $\phi_1^j$  does not necessarily have unit length.

Finally, for both the pseudo-arclength and Moore–Penrose continuation methods one can prove that they converge (with superlinear convergence), provided that  $y_0$  is a regular point and the stepsize is sufficiently small.

### Discretization of BVPs

In this section, *orthogonal collocation* [3,15] is described, a discretization technique to approximate the solution of a generic BVP (3) by a suitable AP. Let  $u$  be at least in the space  $C^1([0, 1], \mathbb{R}^n)$  of continuous differentiable vector-valued functions defined on  $[0, 1]$ . For BVPs the rescaled



**Numerical Bifurcation Analysis, Figure 3**

Pseudo-arclength (a) and Moore–Penrose (b) continuation. Searching a solution in hyperplanes without (a) and with (b) updating the tangent vector. The *open dots* correspond to Newton iterations, *full dots* to points on the curve. The *dotted lines* indicate subspaces where solutions are searched

time  $t = T\tau$  is used, so that the period  $T$  becomes a parameter and in the sequel  $T$  will be addressed as such. Introduce a time mesh  $0 = \tau_0 < \tau_1 < \dots < \tau_N = 1$  and, on each interval  $[\tau_{j-1}, \tau_j]$ , approximate the function  $u$  by a vector-valued polynomial  $\wp_j$  of degree  $m$ ,  $j = 1, \dots, N$ . The polynomials  $\wp_j$  are determined by imposing the ODE in (3) at  $m$  collocation points  $z_{j,i}$ ,  $i = 1, \dots, m$ , i. e.,

$$\dot{\wp}_j(z_{j,i}) = f(\wp_j(z_{j,i}), p), \quad j = 1, \dots, N, i = 1, \dots, m. \quad (10)$$

One usually chooses the so-called Gauss points as the collocation points, the roots of the  $m$ th order Legendre polynomials. Moreover,  $\wp_1(0)$  and  $\wp_N(1)$  must satisfy the boundary conditions and the whole piecewise polynomial must satisfy the integral conditions.

Counting the number of unknowns, the discretization of (3) leads to  $nN$  polynomials, each with  $(m+1)$  degrees of freedom, plus  $n_p$  free parameters, so there are  $nN(m+1) + n_p$  continuation variables. These variables are matched by  $nmN$  collocation equations from (10),  $n(N-1)$  continuity conditions at the mesh points,  $n_b$  boundary conditions, and  $n_i$  integral conditions, for a total of  $nN(m+1) + n_b + n_i - n$  algebraic equations. Thus, in order for these equations to compose an AP, the number of free parameters is generically  $n_p = n_b + n_i - n + 1$  and, typically, one is the period  $T$ .

The collocation method yields high accuracy with superconvergence at the mesh points [15]. The mesh can also be adapted during continuation, for instance to minimize the local discretization error [68]. The equations can be solved efficiently by exploiting the particular sparsity structure of the Jacobi matrix in the Newton iteration. In particular, a few full but essentially smaller systems are solved instead of one sparse but large system. During this process one finds two nonsingular  $(n \times n)$ -submatrices  $M_0$  and  $M_1$  such that  $M_0 u(0) + M_1 u(1) = 0$ , i. e., the monodromy matrix  $M = -M_1^{-1} M_0$  is found as a by-product. In the case of a periodic BVP ( $u(1) = u(0)$ ) the Floquet multipliers are therefore computed at low computational cost.

## Normal Forms and the Center Manifold

Local bifurcation analysis relies on the reduction of the dynamics of system (1) to a lower-dimensional center manifold  $H_0$  near nonhyperbolic equilibria or limit cycles, i. e., when they bifurcate at some critical parameter  $p_0$ . The existence of  $H_0$  follows from the Center Manifold Theorem (CMT), see, e. g., [11], while the reduction principle is shown in [72]. The reduced ODE has the same dimen-

sion as  $H_0$  given by the number  $n_c$  of critical eigenvalues or nontrivial multipliers (counting multiplicity), and is transformed to a normal form. The power of this approach is that the bifurcation scenario of the normal form is preserved in the original system. The normal form for a specific bifurcation is usually studied only up to a finite order, i. e., truncated, and many diagrams for bifurcations with higher codimension are in principle incomplete due to global phenomena, such as connecting orbits. Also,  $H_0$  is not necessarily unique or smooth [75], but fortunately, one can still draw some useful qualitative conclusions.

The codimension (codim) of a bifurcation is the minimal number of parameters needed to encounter the bifurcation and to unfold the corresponding normal form generically. Therefore, in practice, one finds codim 1 phenomena when varying a single parameter, and continues them as curves in two-parameter planes. Codim 2 phenomena are found as isolated points along codim 1 bifurcation curves. Still, codim 2 bifurcations are important as they are the roots of codim 1 bifurcations, in particular of global phenomena. For this reason they are called organizing centers as, around these points in parameter space, one-parameter bifurcation scenarios change. For parameter-dependent systems the center manifold  $H_0$  can be extended to a parameter-dependent invariant manifold  $H(p)$ ,  $H_0 = H(p_0)$ , so that the bifurcation scenario on  $H(p)$  is preserved in the original system for  $\|p - p_0\|$  sufficiently small.

In the following, the normal forms for all codim 1 bifurcations of equilibria and limit cycles are presented and their bifurcation scenarios discussed. Then a general computational method for the approximation, up to a finite order, of the parameter-dependent center manifold  $H(p)$  is presented. The method gives, as a by-product, explicit formulas for the coefficients of a given normal form in terms of the vector field  $f$  of system (1).

## Normal Forms

Bifurcations can be defined by certain algebraic conditions. For instance, an equilibrium is nonhyperbolic if  $\Re(\lambda) = 0$  holds for some eigenvalue. The simplest possibilities are  $\lambda = 0$  (limit point bifurcation or branch point, though the latter is nongeneric, see Sect. “Branch Switching”) and  $\lambda_{1,2} = \pm i\omega_0$ ,  $\omega_0 > 0$  (Hopf bifurcation). Bifurcations of limit cycles appear if some of the nontrivial multipliers cross the unit circle. The three simplest possibilities are  $\mu = 1$  (limit point of cycles),  $\mu = -1$  (period-doubling) or  $\mu_{1,2} = e^{\pm i\theta_0}$ ,  $0 < \theta_0 < \pi$  (Neimark–Sacker).

At the bifurcation ( $p = p_0$ ), linearization of system (1) near the equilibrium  $x_0$  or around a limit cycle, does not

result in any stability information in the center manifold. In this case, nonlinear terms are also necessary to obtain such knowledge. This is provided by the critical normal form coefficients as discussed below. The state variable in the normal form will be denoted by  $w$  and the unfolding parameter by  $\alpha \in \mathbb{R}$ , with  $w = 0$  at  $\alpha = 0$  being a non-hyperbolic equilibrium. Bifurcations are labeled in accordance with the scheme of [38].

### Codimension 1 Bifurcations of Equilibria

**Limit point bifurcation (LP):** The equilibrium has a simple eigenvalue  $\lambda = 0$  and the restriction of (1) to a one-dimensional center manifold can be transformed to the normal form

$$\dot{w} = \alpha + a_{LP}w^2 + O(|w|^3), \quad w \in \mathbb{R}, \quad (11)$$

where generically  $a_{LP} \neq 0$  and  $O$  denotes higher order terms in state variables depending on parameters too. When the unfolding parameter  $\alpha$  crosses the critical value ( $\alpha = 0$ ), two equilibria, one stable and one unstable in the center manifold, collide and disappear. This bifurcation is also called *saddle-node*, *fold* or *tangent* bifurcation. Note that this bifurcation occurs at the folds in Figs. 2 and 3.

**Hopf bifurcation (H):** The equilibrium has a complex pair of eigenvalues  $\lambda_1 = -\lambda_2 = i\omega_0$  and the restriction of (1) to the two-dimensional center manifold is given by

$$\dot{w} = (i\omega_0 + \alpha)w + c_H w^2 \bar{w} + O(|w|^4), \quad w \in \mathbb{C}, \quad (12)$$

where generically the first Lyapunov coefficient  $d_H = \Re(c_H) \neq 0$ . When  $\alpha$  crosses the critical value, a limit cycle is born. It is stable (and present for  $\alpha > 0$ ) if  $d_H < 0$  and unstable if  $d_H > 0$  (and present for  $\alpha < 0$ ). The case  $d_H < 0$  is called supercritical or “soft”, while  $d_H > 0$  is called subcritical or “hard” as there is no (local) attractor left after the bifurcation. This bifurcation is most often called Hopf, but also Poincaré–Andronov–Hopf as this was also known to the first two.

**Codimension 1 Bifurcations of Limit Cycles** Bifurcations of limit cycles are theoretically very well understood using the notion of a Poincaré map. To define this map, choose a  $(n - 1)$ -dimensional smooth cross-section  $\Sigma$  transversal to the cycle and introduce a local coordinate  $z \in \mathbb{R}^{n-1}$  such that  $z = Z(x)$  is defined on  $\Sigma$  and invertible. For example, one chooses a coordinate plane  $x_j = 0$  such that  $f_j(x)|_{x_j=0} \neq 0$ . Let  $x_0(t)$  be the cycle with period  $T$ , so that  $z_0 = Z(x_0(0))$  is the cycle intersection with  $\Sigma$ , where  $z = 0$  can always be assumed without loss of generality. Denote by  $T(z)$  the return time to  $\Sigma$  defined by the flow  $\Phi$  with  $T(z_0) = T$ . Now, the

Poincaré map  $P: \mathbb{R}^{n-1} \rightarrow \mathbb{R}^{n-1}$  maps each point close enough to  $z = 0$  to the next return point on  $\Sigma$ , i.e.,  $P: z \mapsto Z(\Phi^{T(z)}(Z^{-1}(z)))$ . Thus, bifurcations of limit cycles turn into bifurcations of fixed points of the Poincaré map which can be easily described using local bifurcation theory. Moreover, it can be shown that the  $n - 1$  eigenvalues of the linearization  $P_z(0)$  are the nontrivial eigenvalues of the monodromy matrix  $M = \Phi_x^T(x_0(0))$ , which also has a trivial eigenvalue equal to 1 (the vector  $f(x_0(0), p)$ , tangent to the cycle at  $x_0(0)$ , is mapped by  $M$  to itself). The eigenvalues of  $P_z(0)$  are therefore the nontrivial multipliers of the cycle. Although the Poincaré map and its linearization can also be computed numerically through suitably organized simulations (so-called *shooting* techniques [17]), it is better to handle both the cycle multipliers and normal form computations associated to nonhyperbolic cycles using BVPs [9,23,57]. Here, however, the normal forms on a Poincaré section are presented, where  $w = 0$  at  $\alpha = 0$  is the fixed point of the Poincaré map corresponding to a nonhyperbolic limit cycle.

**Limit point of cycles (LPC):** The fixed point has one simple nontrivial multiplier  $\mu = 1$  on the unit circle and the restriction of  $P$  to a one-dimensional center manifold has the form

$$w \mapsto \alpha + w + a_{LPC}w^2 + O(w^3), \quad w \in \mathbb{R},$$

where  $a_{LPC} \neq 0$ . As for the LP bifurcation two fixed points collide and disappear when  $\alpha$  crosses the critical value, provided  $a_{LPC} \neq 0$ . This implies the collision of two limit cycles of the original vector field  $f$ .

**Period-doubling (PD):** The fixed point has one simple multiplier  $\mu = -1$  on the unit circle and the restriction of  $P$  to a one-dimensional center manifold can be transformed to the normal form

$$w \mapsto -(1 + \alpha)w + b_{PD}w^3 + O(w^4), \quad w \in \mathbb{R},$$

where  $b_{PD} \neq 0$ . When the parameter  $\alpha$  crosses the critical value and  $b_{PD} \neq 0$ , a cycle of period 2 for  $P$  bifurcates from the fixed point corresponding to a limit cycle of period  $2T$  for the original system (1). This phenomenon is also called the *flip* bifurcation. If  $b_{PD}$  is positive [negative], the bifurcation is supercritical [subcritical] and the double period cycle is stable [unstable] (and present for  $\alpha > 0$  [ $\alpha < 0$ ]).

**Neimark–Sacker (NS):** The fixed point has simple critical multipliers  $\mu_{1,2} = e^{\pm i\theta_0}$  and no other multipliers on the unit circle. Assume that  $e^{ik\theta_0} \neq 1$  for  $k = 1, 2, 3, 4$ , i.e., there are no strong resonances. Then, the restriction of  $P$  to a two-dimensional center manifold can be transformed to the normal form

$$w \mapsto e^{i\theta(\alpha)}(1 + \alpha)w + c_{NS}w^2 \bar{w} + O(|w|^4), \quad w \in \mathbb{C},$$

where  $c_{\text{NS}}$  is a complex number and  $\theta(0) = \theta_0$ . Provided  $d_{\text{NS}} = \Re(e^{-i\theta_0} c_{\text{NS}}) \neq 0$ , a unique closed invariant curve for  $P$  appears around the fixed point, when  $\alpha$  crosses the critical value. In the original vector field, this corresponds to the appearance of a two-dimensional torus with (quasi-) periodic motion. This bifurcation is also called *secondary Hopf* or *torus* bifurcation. If  $d_{\text{NS}}$  is negative [positive], the bifurcation is supercritical [subcritical] and the invariant curve (torus) is stable [unstable] (and present for  $\alpha > 0$  [ $\alpha < 0$ ]).

### Center Manifolds

Generally speaking, the CMT allows one to restrict the dynamics of (1) to a suspended system

$$\dot{w} = G(w, \alpha), \quad G: \mathbb{R}^{n_c} \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{n_c}, \quad (13)$$

on the center manifold  $H$  and here  $n_p$  is typically 1 or 2 depending on the codimension of the bifurcation. Although the normal forms (13) to which one can restrict the system near nonhyperbolic equilibria and cycles are known, these results are not directly applicable. Thus, efficient numerical algorithms are needed in order to verify the nondegeneracy conditions in the normal forms listed above.

Here, a powerful normalization method due to Iooss and coworkers is reviewed, see [9,14,29,37,55,59]. This method assumes very little a priori information, actually only the type of bifurcation such that the form, i.e., the nonzero coefficients of  $G$ , is known. This fits very well in a numerical bifurcation setting where one computes families of solutions and monitors and detects the occurrence of bifurcations with higher codimension during the continuation.

Without loss of generality it is assumed that  $x_0 = 0$  at the bifurcation point  $p_0 = 0$ . Expand  $f(x, p)$  in Taylor series

$$f(x, p) = Ax + \frac{1}{2}B(x, x) + \frac{1}{6}C(x, x, x) + J_1p + A_1(x, p) + \dots, \quad (14)$$

parametrize, locally to  $(x, p) = (0, 0)$ , the parameter-dependent center manifold by

$$x = H(w, \alpha), \quad H: \mathbb{R}^{n_c} \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}^n, \quad (15)$$

and define a relation  $p = V(\alpha)$  between the original and unfolding parameters. The invariance of the center manifold can be exploited by differentiating this parametrization with respect to time to obtain the so-called *homological equation*

$$f(H(w, \alpha), V(\alpha)) = H_w(w, \alpha)G(w, \alpha). \quad (16)$$

To verify nondegeneracy conditions, only an approximation to the solution of the homological equation is required. To this end,  $G, H$  and  $V$  are expanded in Taylor series:

$$\begin{aligned} G(w, \alpha) &= \sum_{|\mu|+|v|\geq 1} \frac{1}{\mu!v!} g_{\mu v} w^\mu \alpha^v, \\ H(w, \alpha) &= \sum_{|\mu|+|v|\geq 1} \frac{1}{\mu!v!} h_{\mu v} w^\mu \alpha^v, \\ V(\alpha) &= v_{10}\alpha_1 + v_{01}\alpha_2 + O(\|\alpha\|^2), \end{aligned} \quad (17)$$

where  $g_{\mu v}$  are the desired normal form coefficients and  $\mu, v$  are multi-indices. For a multi-index  $\mu$  one has  $\mu = (\mu_1, \mu_2, \dots, \mu_n)$  for nonnegative integers  $\mu_i$ ,  $\mu! = \mu_1! \mu_2! \dots \mu_n!$ ,  $|\mu| = \mu_1 + \mu_2 + \dots + \mu_n$ ,  $\tilde{\mu} \leq \mu$  if  $\tilde{\mu}_i \leq \mu_i$  for all  $i = 1, \dots, n$  and  $w^\mu = w_1^{\mu_1} \dots w_n^{\mu_n}$ . When dealing with just the critical coefficients, i.e.,  $\alpha = 0$ , the index  $v$  is omitted. Substitution of this ansatz into (16) gives a formal power series in  $w$  and  $\alpha$ . As both sides should be equal for all  $w$  and  $\alpha$ , the coefficients of the corresponding powers should be equal. For each vector  $h_{\mu v}$ , (16) gives linear systems of the form

$$L_{\mu v} h_{\mu v} = R_{\mu v}, \quad (18)$$

where  $L_{\mu v} = A - \gamma_{\mu v} I_n$  ( $\gamma_{\mu v}$  is a weighted sum of the critical eigenvalues) and  $R_{\mu v}$  involves known quantities of  $G$  and  $H$  of order less than or equal to  $|\mu| + |v|$ . This leads to an iterative procedure, where, either system (18) is nonsingular, or the required coefficients  $g_{\mu v}$  are obtained by imposing solvability, i.e.,  $R_{\mu v}$  lies in the range of  $L_{\mu v}$  and is therefore orthogonal to the eigenvectors of  $L_{\mu v}^\top$  associated to the zero eigenvalue. In the second case, the solution of (18) is not unique, and one typically selects the  $h_{\mu v}$  without components in the nullspace of  $L_{\mu v}$ . However, the nonuniqueness of the center manifold does not affect qualitative conclusions. The parameter transformation, i.e., the  $v_v$ , is obtained by imposing certain conditions on some normal form coefficients, leading to a solvable system.

One can perform an analogous procedure for the Poincaré map by using a Taylor expansion

$$P(z, p) = Az + \frac{1}{2}B(z, z) + \frac{1}{6}C(z, z, z) + \dots \quad (19)$$

and the homological equation for maps

$$P(H(w, \alpha), V(\alpha)) = H(G(w, \alpha), \alpha). \quad (20)$$

The detailed derivation of the formulas for all codim 1 and 2 cases for equilibria and cycles can be found in [9,55,



**Numerical Bifurcation Analysis, Table 1**

Critical normal form coefficients for generic codim 1 bifurcations of equilibria and fixed points. Here,  $A$ ,  $B$  and  $C$  refer to the expansion (14) for equilibria, while for fixed points they refer to (19)

	Eigenvectors	Critical normal form coefficients
LP	$Av = 0$ $A^T w = 0$	$a_{LP} = \frac{1}{2} w^T B(v, v)$
H	$Av = i\omega_0 v$ $A^T w = -i\omega_0 w$	$c_H = \frac{1}{2} \tilde{w}^T [C(v, v, \tilde{v}) + 2B(v, h_{11}) + B(\tilde{v}, h_{20})]$ $h_{11} = -A^{-1} B(v, \tilde{v}), h_{20} = (2i\omega_0 I_n - A)^{-1} B(v, v)$
LPC	$Av = v$ $A^T w = w$	$a_{LPC} = \frac{1}{2} w^T B(v, v)$
PD	$Av = -v$ $A^T w = -w$	$b_{PD} = \frac{1}{6} w^T [C(v, v, v) + 3B(v, h_2)]$ $h_2 = (I_{n-1} - A)^{-1} B(v, v)$
NS	$Av = e^{i\theta_0} v$ $A^T w = e^{-i\theta_0} w$ $e^{ik\theta_0} \neq 1, k = 1, 2, 3, 4$	$c_{NS} = \frac{1}{2} \tilde{w}^T [C(v, v, \tilde{v}) + 2B(v, h_{11}) + B(\tilde{v}, h_{20})]$ $h_{11} = (I_{n-1} - A)^{-1} B(v, \tilde{v}), h_{20} = (e^{2i\theta_0} I_{n-1} - A)^{-1} B(v, v)$

[56,59,60]. The formulas for the critical normal form coefficients for codim 1 bifurcations are presented in Table 1. Note once more that for limit cycles a numerical more appropriate method exists [57] based on periodic normal forms [47,48].

### Continuation and Detection of Bifurcations

Along a solution branch one generically passes through bifurcation points of higher codimension. To detect such an event, a test function  $\varphi$  is defined, where the event corresponds to a regular zero. If at two consecutive points  $y_{k-1}, y_k$  along the branch the test function changes sign, i. e.,  $\varphi(y_k)\varphi(y_{k-1}) < 0$ , then the zero can be located more precisely. Usually, a one-dimensional secant method is used to find such a point. Now, if system (1) has a bifurcation at  $y_0 = (x_0, p_0)$ , then there is generically a curve  $y = y(s)$  where the system displays this bifurcation. In order to find this curve, one starts with a known point  $y_0$  and formulates a defining system and then continue that solution in one extra free parameter.

### Test Functions for Codimension 1 Bifurcations

An equilibrium may lose stability through a limit point, a Hopf bifurcation or in a branch point. At a limit or branch point bifurcation the Jacobi matrix  $A = f_x(x_0, p_0)$  has an algebraically simple eigenvalue  $\lambda = 0$  (see Sect. “Branch Switching” for branch points), while at a Hopf point there is a pair of complex conjugate eigenvalues  $\lambda = \pm i\omega_0, \omega_0 \neq 0$  and only one such pair.

The simplest way of detecting the passage through a bifurcation during continuation, is to monitor the eigenval-

ues of the Jacobi matrix. For large systems and stiff problems this is prohibitive as it is numerically expensive and not always accurate. Instead, one can base test functions on determinants.

**Test Functions for Limit Point Bifurcations** Along an equilibrium curve the product of the eigenvalues changes sign at a limit point. Recall that the determinant of  $A$  is the product of its eigenvalues. Therefore, the following test function can be computed

$$\varphi_{LP} = \det(f_x(x, p)) \quad (21)$$

without computing the eigenvalues explicitly.

For the LP bifurcation the pseudo-arclength or Moore–Penrose continuation methods provide an excellent test function as a by-product of the continuation. Note that while passing through the fold, the last component of the tangent vector  $\phi$  changes sign as the continuation direction in the parameter reverses. The test function is therefore defined as

$$\varphi_{LP} = \phi_{n+1} \cdot \quad (22)$$

**Test Functions for Hopf Bifurcations** Denote the eigenvalues of  $A$  by  $\lambda_i(x, p)$ ,  $i = 1 \dots, n$  and consider the following product

$$\varphi_H = \prod_{i < j} (\lambda_i(x, p) + \lambda_j(x, p)) \cdot$$

It can be shown that this product has a regular zero at a simple Hopf point [9], but it should be checked that this

zero corresponds to an imaginary pair and not to the neutral saddle case  $\lambda_i = -\lambda_j$ ,  $\lambda_i \in \mathbb{R}$ .

Also here one can compute this product without explicit computation of the eigenvalues using the bi-alternate product [34,42,45,56]. The bi-alternate product of two  $(n \times n)$ -matrices  $A$  and  $B$ , denoted by  $A \odot B$ , is a  $(m \times m)$ -matrix  $C$  ( $m = n(n-1)/2$ ) with row index  $(i, j)$  and column index  $(k, l)$  and elements

$$C_{(i,j)(k,l)} = \frac{1}{2} \left\{ \begin{vmatrix} a_{ik} & a_{il} \\ b_{jk} & b_{jl} \end{vmatrix} + \begin{vmatrix} b_{ik} & b_{il} \\ a_{jk} & a_{jl} \end{vmatrix} \right\}$$

where  $i = 2, 3, \dots, n, j = 1, 2, \dots, i-1,$   
 $k = 2, 3, \dots, n, l = 1, 2, \dots, k-1.$

Let  $A$  be an  $n \times n$ -matrix with eigenvalues  $\lambda_1, \dots, \lambda_n$ , then [73]

- $A \odot A$  has eigenvalues  $\lambda_i \lambda_j$ ,
- $2A \odot I_n$  has eigenvalues  $\lambda_i + \lambda_j$ .

The test function can now be expressed as

$$\varphi_H = \det(2f_x(x, p) \odot I_n). \quad (23)$$

For higher dimensional systems, this matrix becomes very large and one should use precondition or subspace methods, see [36].

### Test Functions for Codimension 1 Cycle Bifurcations

Recall that the nontrivial multipliers  $\mu_1, \dots, \mu_{n-1}$  determine the stability of the cycle and can be efficiently computed as the nontrivial multipliers of the monodromy matrix  $M$ , see Sect. “Discretization of BVPs”. Now the following two sets of test functions can be used to detect LPC, PD and NS bifurcations

$$\begin{aligned} \varphi_{\text{LPC}} &= \prod_{i=1}^{n-1} (\mu_i - 1), & \varphi_{\text{LPC}} &= \phi_p, \\ \varphi_{\text{PD}} &= \prod_{i=1}^{n-1} (\mu_i + 1), & \varphi_{\text{PD}} &= \det(M + I_n), \\ \varphi_{\text{NS}} &= \prod_{1 \leq i < j}^{n-1} (\mu_i \mu_j - 1), & \varphi_{\text{NS}} &= \det(M \odot M - I_{n(n-1)/2}). \end{aligned}$$

where  $\phi_p$  denotes the parameter component of the tangent vector similar to (22). It should also be checked that a zero of  $\varphi_{\text{NS}}$  corresponds to nonreal multipliers  $e^{i\theta_0}$ , similar to the test function to detect the Hopf bifurcation.

There are alternatives for these test functions. One can define bordered systems using the monodromy matrix [9,42] or a BVP formulation [23].

### Defining Systems for Codimension 1 Bifurcations of Equilibria

To compute curves of codim 1 equilibria bifurcations, first a defining system of the form (2) needs to be formulated to define the bifurcation curve, and then a second parameter for the continuation must be freed, so that now  $p \in \mathbb{R}^2$ . This is done by adding to the equilibrium equation  $f(x, p) = 0$  appropriate equations that characterize the bifurcation.

Defining systems come in two flavors, *fully* (also standard) and *minimally* augmented systems. The first computes all relevant eigenspaces, while the latter exploits the rank deficiency of the Jacobi matrix and adds only a few strategic equations to regularize the continuation problem. The evaluation of such equations requires the eigenspaces, but these can be computed separately. As the names suggest, the difference is in the dimension of the defining system leading to differently sized problems. In particular, the advantage of minimally augmented systems is that of solving several smaller linear problems, instead of a big one, which is known to be better in terms of both accuracy and computational time. For small phase dimension  $n$  there is little difference in computational effort. Both minimally and fully extended defining systems for both limit point and Hopf bifurcations are presented. The regularity of these systems is also known, e. g., see [42].

**Defining Systems for Limit Point Bifurcations** The first defining system is minimally extended adding the test function (21)

$$\begin{cases} f(x, p) = 0, \\ \det(f_x(x, p)) = 0. \end{cases} \quad (24)$$

This system consists of  $n+1$  equations in  $n+2$  unknowns  $(x, p)$ . One problem is that the computation of the determinant can lose accuracy for large systems. This can be avoided in two ways, by augmenting the system with the eigenspaces or using a bordering technique.

Fully extended systems include the eigenvectors and for a LP bifurcation this leads to

$$\begin{cases} f(x, p) = 0, \\ f_x(x, p)v = 0, \\ v_0^\top v - 1 = 0, \end{cases} \quad (25)$$

where  $v_0$  is a vector not orthogonal to  $\mathcal{N}(f_x(x, p))$ . This system consists of  $2n+1$  equations in  $2n+2$  unknowns  $(x, p, v)$ .

The bordering technique uses the bordering lemma [41]. Let  $A \in \mathbb{R}^{n \times n}$  be a singular matrix and let

$B, C \in \mathbb{R}^{n \times m}$  such that the system

$$\begin{pmatrix} A & B \\ C^\top & 0_m \end{pmatrix} \begin{pmatrix} V \\ g \end{pmatrix} = \begin{pmatrix} 0_{n \times m} \\ I_m \end{pmatrix} \quad (26)$$

is nonsingular ( $V \in \mathbb{R}^{n \times m}$ ,  $g \in \mathbb{R}^{m \times m}$ ). Typically  $B$  and  $C$  are associated to the eigenspaces of  $A^\top$  and  $A$  corresponding to the zero eigenvalue, respectively or, during continuation, approximated by their values computed at the previous point along the branch. It follows from the bordering lemma that  $A$  has rank deficiency  $m$  if and only if  $g$  has rank deficiency  $m$ .

With  $A = f_x(x, p)$  and  $m = 1$ , one has  $g = 0$  if and only if  $\det(f_x(x, p)) = 0$ . A modified and minimally extended system for limit points is thus given by

$$\begin{cases} f(x, p) = 0, \\ g(x, p) = 0, \end{cases}$$

where  $g$  is defined by (26) with  $A^\top B = AC = 0$  at a previously computed point. During the continuation the derivatives of  $g$  w.r.t. to  $x$  and  $p$  are needed. They can either be approximated by finite differences, or explicitly (and efficiently) obtained from the second-derivatives of the vector field  $f$ , see [42].

**Defining Systems for Hopf Bifurcations** Defining systems for Hopf bifurcations are formulated analogously to the LP case. Adding the test function (23) creates a minimally extended system

$$\begin{cases} f(x, p) = 0, \\ \det(2f_x(x, p) \odot I_n) = 0, \end{cases} \quad (27)$$

while the fully extended system is given by

$$\begin{cases} f(x, p) = 0, \\ f_x(x, p)v_1 + \omega v_2 = 0, \\ f_x(x, p)v_2 - \omega v_1 = 0, \\ w_1^\top v_1 + w_2^\top v_2 - 1 = 0, \\ w_1^\top v_2 - w_2^\top v_1 = 0, \end{cases} \quad (28)$$

where  $w = w_1 + iw_2$  is not orthogonal to the eigenvector  $v = v_1 + iv_2$  corresponding to the eigenvalue  $i\omega$ . The vector  $w = v^{k-1}$  computed at the previous point is a suitable choice during continuation. System (28) is expressed using real variables and has  $3n + 2$  equations for  $3n + 3$  unknowns  $(x, p, v_1, v_2, \omega)$ .

A reduced defining system can be obtained from (28) by noting that the matrix  $f_x(x, p)^2 + \kappa I_n$  has rank deficiency two at a Hopf bifurcation point with  $\kappa = \omega^2$  [67].

An alternative to (28) is now formulated as

$$\begin{cases} f(x, p) = 0, \\ [f_x(x, p)^2 + \kappa I_n]v = 0, \\ v^\top v - 1 = 0, \\ w^\top v = 0, \end{cases} \quad (29)$$

where  $w$  is not orthogonal to the two-dimensional real eigenspace of the eigenvalues  $\pm i\omega$ . It has  $2n + 2$  equations for  $2n + 3$  unknowns  $(x, p, v, \kappa)$ . However,  $w$  needs to be updated during continuation, e.g., as the solution of  $([f_x(x, p)^2 + \kappa I_n]^\top w, v^\top w) = (0, 0)$  computed at the previous continuation point.

A further reduction is obtained exploiting the rank deficiency. Consider the system

$$\begin{pmatrix} f_x(x, p)^2 + \kappa I_n & B \\ C^\top & 0_2 \end{pmatrix} \begin{pmatrix} V \\ g \end{pmatrix} = \begin{pmatrix} 0_{n \times 2} \\ I_2 \end{pmatrix}$$

and it follows from the bordering lemma that  $g$  vanishes at Hopf points and any two components of  $g$ , e.g.,  $g_{11}$  and  $g_{22}$ , see [42], can be taken to augment Eq. (2) to obtain the following minimally augmented system

$$\begin{cases} f(x, p) = 0, \\ g_{11} = 0, \\ g_{22} = 0, \end{cases} \quad (30)$$

which has  $n + 2$  equations for  $n + 3$  unknowns  $(x, p, \kappa)$ .

### Defining Systems for Codimension 1 Bifurcations of Limit Cycles

In principle, to study bifurcations of limit cycles one can compute numerically the Poincaré map and study bifurcations of fixed points. If system (1) is not stiff, then the Poincaré map and its derivatives may be obtained with satisfactory accuracy. In many cases, however, continuation using BVP formulations is much more efficient.

Suppose a cycle  $x$  bifurcates at  $p = p_0$ , then the BVP (4) defining the limit cycle must be augmented with suitable extra functions. As for codim 1 branches of equilibria, one can define either fully extended systems by including the relevant eigenfunctions in the computation [9], or minimally extended systems using bordered BVPs [23,39]. The regularity of these defining systems is also discussed in these references. Since the discretization of the cycle leads to large APs, here the minimally extended approach can lead to faster results even though some more algebra is involved, see the comparison in [57]. Below only the equations are presented, which are added to the defining system (4) for the continuation of limit cycles.

**Fully Extended Systems** The following equations can be used to augment (4) and continue codim 1 bifurcations of limit cycles. The eigenfunctions  $v$  need to be discretized in a similar way to that in Sect. “Discretization of BVPs”. The previous cycle  $x^{k-1}$  and eigenfunction  $v^{k-1}$  are assumed to be known.

LPC: For the *limit point of cycles* bifurcation, the BVP (4) is augmented with the equations

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) - \sigma f(x(\tau), p) = 0, \\ v(1) - v(0) = 0, \\ \int_0^1 v^\top(\tau) \dot{x}^{k-1}(\tau) d\tau = 0, \\ \int_0^1 v^\top(\tau) v(\tau)^{k-1} d\tau + \sigma \sigma^{k-1} = 1, \end{cases} \quad (31)$$

for the variables  $(x, p, T, v, \sigma)$ . Note that

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) - Tf_p(x(\tau), p)q \\ - \sigma f(x(\tau), p) = 0, \\ v(1) - v(0) = 0, \\ \int_0^1 v^\top(\tau) \dot{x}^{k-1}(\tau) d\tau = 0, \\ \int_0^1 v^\top(\tau) v(\tau)^{k-1} d\tau + q^{k-1}q + \sigma^{k-1}\sigma = 1, \end{cases}$$

defines the tangent vector  $\phi = (v, q, \sigma)$  to the solution branch, so that (31) simply imposes  $q = 0$ , i.e., the limit point. Together with (4), they compose a BVP with  $2n$  ODEs,  $2n$  boundary conditions, and 2 integral conditions, i.e.,  $n_p = 2n + 2 - 2n + 1 = 3$ , namely  $T$  and two free parameters. Similar dimensional considerations hold for the PD and NS cases below.

PD: For the *period-doubling* bifurcation, the extra equations augmenting (4) are

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) = 0, \\ v(1) + v(0) = 0, \\ \int_0^1 v^\top(\tau) v^{k-1}(\tau) d\tau = 1, \end{cases} \quad (32)$$

for the variables  $(x, p, T, v)$ . Here  $v$  is the eigenfunction of the linearized ODE associated with the multiplier  $\mu = -1$ . In fact, the second equation in (32) imposes  $v(1) = Mv(0) = -v(0)$ , where  $M$  is the monodromy matrix, while the third equation scales the eigenfunction against the previous continuation point.

NS: For the *Neimark–Sacker* bifurcation, the BVP (4) is augmented with the equations

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) = 0, \\ v(1) - e^{i\theta}v(0) = 0, \\ \int_0^1 \bar{v}^\top(\tau) v^{k-1}(\tau) d\tau = 1, \end{cases} \quad (33)$$

for the variables  $(x, p, T, v, \theta)$  with  $v \in C^1([0, 1], \mathbb{C}^n)$ . Here  $v$  is the eigenfunction of the linearized ODE associated with the multiplier  $\mu = e^{i\theta}$ . Of course, the real formulation should be used in practice.

**Minimally Extended Systems** For limit cycle continuation the discretization of the fully extended BVP (4) with (31), (32) or (33) may lead to large APs to be solved. In [39] a minimally extended formulation is proposed to augmenting (4) with a function  $g$  with only a few components. The corresponding function  $g$  is defined using bordered systems.

LPC: For this bifurcation, one uses suitable bordering functions  $v_1, w_1$  and vectors  $v_2, w_2, w_3$  such that the following system linear in  $(v, \sigma, g)$  is regular

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v - f(x(\tau), p)\sigma + w_1g = 0, \\ v(1) - v(0) + w_2g = 0, \\ \int_0^1 f(x(\tau), p)^\top v(\tau) d\tau + w_3g = 0, \\ \int_0^1 v_1^\top v(\tau) d\tau + v_2\sigma = 1. \end{cases} \quad (34)$$

The function  $g = g(x, T, p)$  vanishes at a LPC point. The bordering functions  $v_1, w_1$  and vectors  $v_2, w_2, w_3$  can be updated to keep (34) nonsingular, in particular,  $v_1 = v^{k-1}$  and  $v_2 = \sigma^{k-1}$  from the previously computed point are used. It is convenient to introduce the Dirac operator  $\delta_i f = f(i)$  and the integral operator  $\text{Int}_{v(\cdot)} f = \int_0^1 v(\tau)^\top f(\tau) d\tau$  and to rewrite (34) in operator form

$$\begin{pmatrix} D - Tf_x(x(\cdot), p) & -f(x(\cdot), p) & w_1 \\ \delta_0 - \delta_1 & 0 & w_2 \\ \text{Int}_{f(x(\cdot), p)} & 0 & w_3 \\ \text{Int}_{v_1(\cdot)} & v_2 & 0 \end{pmatrix} \begin{pmatrix} v \\ \sigma \\ g \end{pmatrix} = \begin{pmatrix} c0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \quad (35)$$

PD: The same notation as for the minimally extended LPC defining system is used and suitable bordering functions  $v_1, w_1$  and vector  $w_2$  are chosen such that the following system is regular

$$\begin{pmatrix} D - Tf_x(x(\cdot), p) & w_1 \\ \delta_0 + \delta_1 & w_2 \\ \text{Int}_{v_1(\cdot)} & 0 \end{pmatrix} \begin{pmatrix} v \\ g \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \quad (36)$$

At a PD bifurcation  $g(x, T, p)$  defined by (36) vanishes.

NS: Let  $\hat{\kappa} = \cos(\theta)$  denote the real part of the non-hyperbolic multiplier and choose bordering functions  $v_1, v_2, w_{11}, w_{12}$  and vectors  $w_{21}, w_{22}$  such that the following system is nonsingular and defines the four components of  $g$

$$\begin{pmatrix} D - Tf_x(x(\cdot), p) & w_{11} & w_{12} \\ \delta_2 - 2\hat{\kappa}\delta_1 + \delta_0 & w_{21} & w_{22} \\ \text{Int}_{v_1(\cdot)} & 0 & 0 \\ \text{Int}_{v_2(\cdot)} & 0 & 0 \end{pmatrix} \begin{pmatrix} v \\ g \end{pmatrix} = \begin{pmatrix} 0_{n \times 2} \\ I_2 \end{pmatrix}. \quad (37)$$

At a NS bifurcation the four components of  $g(x, T, p)$  defined by (37) vanish, and, similar to the Hopf bifurcation, the BVP (4) can be augmented with any two components of  $g$ .

### Test Functions for Codimension 2 Bifurcations

During the continuation of codim 1 branches, one meets generically codim 2 bifurcations. Some of which arise through extra instabilities in the linear terms, while other codim 2 bifurcations are defined through degeneracies in the normal form coefficients. For equilibria, codim 2 bifurcations of the first type are the Bogdanov–Takens (BT, two zero eigenvalues with only one associated eigenvector), the zero-Hopf (ZH, also called Gavrilov-Guckenheimer, a simple zero eigenvalue and a simple imaginary pair), and the double Hopf (HH, two distinct imaginary pairs), while higher order degeneracies lead to cusp (CP,  $a_{LP} = 0$  in the normal form (11)) or generalized Hopf (GH,  $d_H = 0$

in the normal form (12), also called Bautin or degenerate Hopf). For cycles, there are strong resonances (R1–R4), fold-flip (LPPD), fold-Neimark–Sacker (LPNS), flip-Neimark–Sacker (PDNS), and double Neimark–Sacker (NSNS) among those involving linear terms, while higher order degeneracies lead to cusp (CP), degenerate flip (GPD), and Chenciner (CH) bifurcations. Naturally the normal form coefficients, see [9,56], are a suitable choice for the corresponding test functions. In Tables 2 and 3, test functions are given which are defined along the corresponding codim 1 branches of equilibrium and limit cycle bifurcations, respectively. The functions refer to the corresponding defining system and to Table 1. Upon detecting and locating a zero of a test function it may be necessary to check that a bifurcation is really involved, similar to the Hopf case where neutral saddles are excluded. For details about the dynamics and the bifurcation diagrams at codim 2 points, see [2,44,56].

### Branch Switching

This section considers points in the continuation space from which several solution branches of interest, with the same codimension, emanate. At these points suitable “branch switching” procedures are required to switch from one solution branch to another. First, the transversal intersection of two solution branches of the same continuation problem is considered, which occurs at so-called Branch Points (BP) (also called “singular” or “transcritical” bifurcation points). Branch points are nongeneric, in the sense that arbitrarily small perturbations of  $F$  in (2) turn the intersection into two separated branches, which come close to the “ghost” of the (disappeared) intersection.

#### Numerical Bifurcation Analysis, Table 2

Test functions along limit point and Hopf bifurcation curves. The matrix  $A_c$  for the test function of the double Hopf bifurcation can be obtained as the orthogonal complement in  $\mathbb{R}^n$  of the Jacobi matrix  $A$  w.r.t. the two-dimensional eigenspace associated with the computed branch of Hopf bifurcations

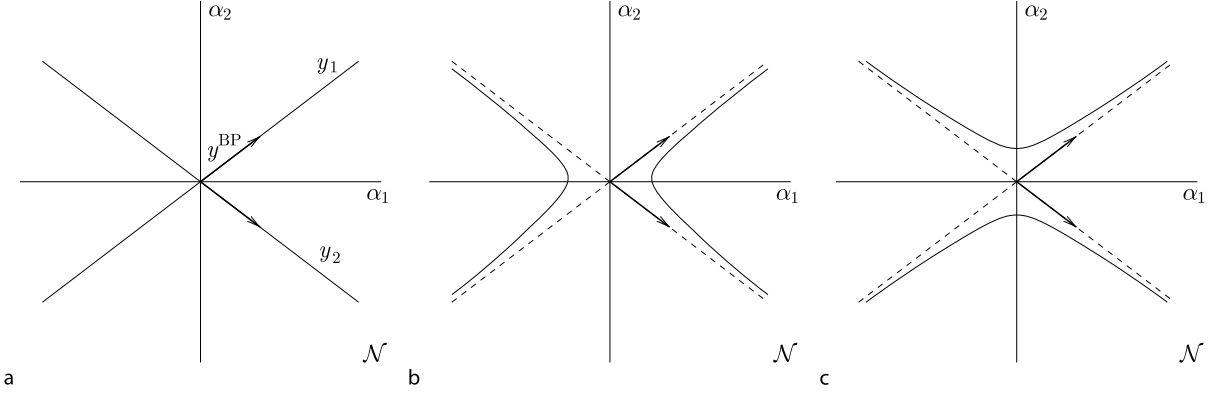
	Label	LP	H
cusp	CP	$a_{LP}$	
generalized Hopf	GH		$d_H$
Bogdanov–Takens	BT	$w_{LP}^T v_{LP}$	$\kappa$
zero-Hopf	ZH	$\varphi_H$	$\varphi_{LP}$
double Hopf	HH		$\det(2A_c \odot I_{n-2})$

#### Numerical Bifurcation Analysis, Table 3

Test functions along LPC, PD and NS bifurcation curves. The matrix  $M_c$  along the Neimark–Sacker bifurcation curve is defined similarly as  $A_c$  in Table 2 along the Hopf bifurcation as the orthogonal complement of the monodromy matrix  $M$  w.r.t. the two-dimensional eigenspace associated with the computed branch of Neimark–Sacker bifurcations

	Label	LPC	PD	NS
cusp	CP	$a_{LPC}$		
degenerate flip	GPD		$b_{PD}$	
Chenciner	CH			$d_{NS}$
resonance 1:1	R1	$w_{LP}^T v_{LP}$		$\hat{\kappa} - 1$
resonance 1:2	R2		$w_{PD}^T v_{PD}$	$\hat{\kappa} + 1$
resonance 1:3	R3			$\hat{\kappa} + \frac{1}{2}$
resonance 1:4	R4			$\hat{\kappa}$
fold-flip	LPPD	$\varphi_{PD}$	$\varphi_{LP}$	
fold-Neimark–Sacker	LPNS	$\varphi_{NS}$		$\varphi_{LP}$
flip-Neimark–Sacker	PDNS		$\varphi_{NS}$	$\varphi_{PD}$
double Neimark–Sacker	NSNS			$\det(M_c \odot M_c - I_{(n-2)(n-3)/2})$





**Numerical Bifurcation Analysis, Figure 4**

(a) Projection of two solution branches of (2), intersecting at  $y^{\text{BP}}$ , on the null-space of  $F_y(y^{\text{BP}})$  ( $\mathcal{N}$ ), close to  $y^{\text{BP}}$  (planar representation in coordinates  $(\alpha_1, \alpha_2)$  with respect to a given basis). The two solution branches are approximated by the straight lines in  $\mathcal{N}$  spanned by their tangent vectors at  $y^{\text{BP}}$  (thick vectors). (b) and (c) Projection on  $\mathcal{N}$ , close to  $y^{\text{BP}}$ , of the two solution branches of the perturbed problem (see (43) for  $b > 0$  and  $b < 0$ , respectively)

tion but then fold (LP) and leave as if they follow the other branch (see Fig. 4).

BPs, however, are very common in applications due to particular symmetries of the continuation problem, like reflections in state space, conserved quantities or the presence of trivial solutions. This is why BP detection and continuation recently received attention [16,25,28]. Then, the switch from a codim 0 solution branch to that of a different continuation problem at codim 1 bifurcations is examined. In particular, the equilibrium-to-cycle switch at a Hopf bifurcation and the period-1-to-period-2 cycle switch at a flip bifurcation are discussed. Finally, various switches between codim 1 solution branches of different continuation problems at codim 2 bifurcations are addressed.

### Branch Switching at Simple Branch Points

Simple BPs are points  $y^{\text{BP}} = (u^{\text{BP}}, p^{\text{BP}})$ , encountered along a solution branch of (2), at which the nullspace  $\mathcal{N}(F_y(y))$  of  $F_y(y)$  is two-dimensional, i. e., the nullspace is spanned by two independent vectors  $\phi_1, \phi_2 \in Y$ , with  $\phi_i^\top \phi_i = 1$ ,  $i = 1, 2$ . Generically, two solution branches of (2) pass through  $y^{\text{BP}}$ , with transversal tangent vectors given by suitable combinations of  $\phi_1$  and  $\phi_2$ . In the following, only the case of the AP (2), i. e.,  $Y = \mathbb{R}^{n+1}$  ( $u \in \mathbb{R}^n$  and  $p \in \mathbb{R}$ ) and  $F: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$  is considered. Similar considerations hold for the BVP (3) (see [16] for details), though, loosely speaking, results for APs can be applied to BVPs after time discretization.

A BP is not a regular point, since  $\text{rank}(F_y(y^{\text{BP}})) = n - 1$ . Distinguishing between state and parameters, there are two possibilities

$$\begin{cases} \text{(i)} & \dim \mathcal{N}(F_u(y^{\text{BP}})) = 1, F_p(y^{\text{BP}}) \in \mathcal{R}(F_u(y^{\text{BP}})) \\ & \implies \phi_1 = (v_1, 0), \phi_2 = (v_2, q_2), \\ \text{(ii)} & \dim \mathcal{N}(F_u(y^{\text{BP}})) = 2, F_p(y^{\text{BP}}) \notin \mathcal{R}(F_u(y^{\text{BP}})) \\ & \implies \phi_1 = (v_1, 0), \phi_2 = (v_2, 0), \end{cases} \quad (38)$$

for suitably chosen  $v_1, v_2, q_2$ . In particular, in the first case,  $v_1$  spans the nullspace of  $F_u(y^{\text{BP}})$  and  $\phi_2$  is determined by solving  $(F_u(y^{\text{BP}})v_2 + F_p(y^{\text{BP}})q_2, v_1^\top v_2, v_2^\top v_2 + q_2^2) = (0, 0, 1)$ .

BPs can be detected by means of the following test function

$$\varphi_{\text{BP}} = \det \begin{pmatrix} F_y(y) \\ \phi^\top \end{pmatrix}, \quad (39)$$

where  $\phi$  is the tangent vector to the solution branch during continuation, which indeed vanishes when (2) admits a second independent tangent vector. Note from (38) that test function (21) also vanishes at BPs, so that test function (22) is more appropriate for LPs.

The vectors tangent to the two solution branches intersecting at  $y^{\text{BP}}$  can be computed as follows. Parametrize one of the two solution branches by a scalar coordinate  $s$ , e. g., the arclength, so that  $y(s)$  and  $y_s(s)$  denote the branch and its tangent vector locally to  $y(0) = y^{\text{BP}}$ . Then,  $F(y(s))$  is identically equal to zero, so taking twice

the derivative w.r.t.  $s$  one obtains  $F_{yy}(y(s))[y_s(s), y_s(s)] + F_y(y(s))y_{ss}(s) = 0$ , which at  $y^{\text{BP}}$  reads

$$F_{yy}(y^{\text{BP}})[y_s(0), y_s(0)] + F_y(y^{\text{BP}})y_{ss}(0) = 0, \quad (40)$$

with  $y_s(0) = \alpha_1\phi_1 + \alpha_2\phi_2$ . Let  $\psi \in \mathbb{R}^n$  span the nullspace of  $F_y(y^{\text{BP}})^\top$  with  $\psi^\top \psi = 1$ . Since the range of  $F_y(y^{\text{BP}})$  is orthogonal to the nullspace of  $F_y(y^{\text{BP}})^\top$ , one can eliminate  $y_{ss}(0)$  in (40) by left-multiplying both sides by  $\psi^\top$ , thus obtaining

$$\psi^\top F_{yy}(y^{\text{BP}})[\alpha_1\phi_1 + \alpha_2\phi_2, \alpha_1\phi_1 + \alpha_2\phi_2] = 0. \quad (41)$$

Equation (41) is called the *algebraic branching equation* [50] and is often written as

$$c_{11}\alpha_1^2 + 2c_{12}\alpha_1\alpha_2 + c_{22}\alpha_2^2 = 0, \quad (42)$$

with  $c_{ij} = \psi^\top F_{yy}(y^{\text{BP}})[\phi_i, \phi_j]$ ,  $i, j = 1, 2$ . At BP detection, the discriminant  $c_{11}c_{22} - c_{12}^2$  is generically negative (otherwise the BP would be an isolated solution point of (2)), so that two distinct pairs  $(\alpha_1, \alpha_2)$  and  $(\tilde{\alpha}_1, \tilde{\alpha}_2)$ , uniquely defined up to scaling, solve (42) and give the directions of the two emanating branches.

Once the two directions are known, one can easily perform branch switching by an initial prediction from  $y^{\text{BP}}$  along the desired direction. This, however, requires the second-order derivatives of  $F$  w.r.t. all continuation variables. Though good approximations can often be achieved by finite differences, an alternative and computationally cheap prediction can be taken in the nullspace of  $F_y(y^{\text{BP}})$  along the direction orthogonal to  $y_s(0)$ . The vector  $y_s(s)$  is in fact known at each point during the continuation of the solution branch up to BP detection, so that the cheap prediction for the other branch spans the (one-dimensional) nullspace of

$$\begin{bmatrix} F_y(y^{\text{BP}}) \\ y_s(0)^\top \end{bmatrix}.$$

### Branch Point Continuation

**Generic Problems** Several defining systems have been proposed for BP continuation, see [16,25,28,63,64,65]. Among fully extended formulations, the most compact one characterizes BPs as points at which the range of  $F_y(y)$  has rank defect 1, i. e., the nullspace of  $F_y(y)^\top$  is one-dimensional. BP continuation is therefore defined by

$$\begin{cases} F(u, p) = 0, \\ F_u(u, p)^\top \psi = 0, \\ F_p(u, p)^\top \psi = 0, \\ \psi^\top \psi - 1 = 0. \end{cases}$$

Counting equations,  $2n + 2$ , and variables  $u, \psi \in \mathbb{R}^n$ ,  $p \in \mathbb{R}$ , i. e.,  $2n + 1$  scalar variables, it follows that two extra parameters generically need to be freed. In other words, BPs are codim 2 bifurcations, which are not expected along generic solution branches of (2).

**Non-generic Problems** BP continuation can be performed in a single extra free parameter for nongeneric problems characterized by symmetries that persist for all parameter values. In such cases, the continuation problem (2) is perturbed into

$$F(y) + bu_b = 0, \quad (43)$$

where  $b \in \mathbb{R}$  and  $u_b \in \mathbb{R}^n$  are new variables of the defining system. The idea is that  $u_b$  “breaks the symmetry”, in the sense that problem (43) has no BP for small  $b \neq 0$ , and BP continuation can be performed in two extra free parameters, one of which,  $b$ , remains zero during the continuation. The choice of  $u_b$  is not trivial. Geometrically,  $u_b$  must be such that small values of  $b$  perturb Fig. 4a into Fig. 4b, say for  $b > 0$ , and into Fig. 4c for  $b < 0$ . It turns out (see, e. g., [16]) that  $u_b = \psi$  is a good choice, i. e., perturbations not in the range of  $F_y(y^{\text{BP}})$  break the symmetry, since close to the BP, they must be balanced by the nonlinear terms of the expansion of  $F$  in (43), and this implies significant deviations of the perturbed solution branch  $y$  from the unperturbed  $y(s)$ .

BP continuation for nongeneric problems is therefore defined by

$$\begin{cases} F(x, p) + b\psi = 0, \\ F_x(x, p)^\top \psi = 0, \\ F_p(x, p)^\top \psi = 0, \\ \psi^\top \psi - 1 = 0. \end{cases} \quad (44)$$

This defining system is also useful for accurately computing BPs. In fact, the basin of convergence of the Newton iterations in (8) or (9) shrinks at BPs (recall  $F_y(y)$  does not have full rank at BPs), while system (44), in the  $2n + 2$  variables  $(u, p, b, \psi)$ , has a unique solution  $(u^{\text{BP}}, p^{\text{BP}}, 0, \psi)$  close to the BP. Thus, when the BP test function (39) changes sign along a solution branch of (2), Newton corrections can be applied to (44), starting from the best possible prediction, i. e., with  $b = 0$  and  $\psi$  as the eigenvector of  $F_u(u, p)^\top$  associated with the real eigenvalue closest to zero.

**Minimally Extended Formulation** A minimally extended defining system for BP continuation requires two scalar conditions,  $g_1(u, p) = 0$ ,  $g_2(u, p) = 0$ , to be added to the unperturbed or perturbed problem (2) or (43)

for generic and nongeneric problems, respectively. These functions  $g_1$  and  $g_2$  are defined in [28] by solving

$$\begin{cases} F_y(y)\phi_1 + g_1\psi^{k-1} = 0, \\ F_y(y)\phi_2 + g_2\psi^{k-1} = 0, \\ (\phi_1^{k-1})^\top \phi_1 - 1 = 0, \\ (\phi_2^{k-1})^\top \phi_2 - 1 = 0, \\ (\phi_1^{k-1})^\top \phi_2 = 0, \\ (\phi_2^{k-1})^\top \phi_1 = 0, \end{cases}$$

in the unknowns  $\phi_1, \phi_2, g_1, g_2$ , while  $\psi$  is updated by solving

$$\begin{cases} F_y(y)^\top \psi + g_1\phi_1 + g_2\phi_2 = 0, \\ \psi^\top \psi - 1 = 0, \end{cases}$$

in the unknowns  $\psi, g_1, g_2$ , after each Newton convergence.

### Branch Switching at Hopf Points

At a Hopf bifurcation point  $y^H = (x^H, p^H)$ , one typically wants to start the continuation of the emanating branch of limit cycles. For this, one might think of using the branch switching procedure described above to switch from a constant to a periodic solution branch of the limit cycle BVP (4). Unfortunately,  $y^H$  is not a simple BP for problem (4), since the period  $T$  is undetermined along the constant solution branch, so that, formally, an infinite number of branches emanate from  $y^H$ . Thus, a prediction in the proper direction, i.e., along the vector  $\phi = (v, q)$  tangent to the periodic solution branch, is required.

Let  $y(s)$  represent the periodic solution branch, with  $y(0) = y^H$ . Then,  $x$  and  $v$  are period-1 vector-valued functions in  $C^1([0, 1], \mathbb{R}^n)$ ,  $p \in \mathbb{R}$  and  $T$  are the free parameters, and  $q = (p_s, T_s)$ . The Hopf bifurcation theorem [56] ensures that  $p_s = T_s = 0$  and that  $v$  is the unit-length solution of the linearized, time-independent equation  $\dot{v} = T(0)f_x(x^H, p^H)v$ , i.e.,  $v(\tau) = \sin(2\pi\tau)w_r + \cos(2\pi\tau)w_i$ , where  $w = w_r + iw_i$  ( $w_r^\top w_r + w_i^\top w_i = 1, w_r^\top w_i = 0$ ) is the complex eigenvector of  $f_x(x^H, p^H)$  associated to the eigenvalue  $i\omega$ ,  $\omega = 2\pi/T(0)$ .

The periodic solution branch of the limit cycle BVP (4) can therefore be followed, provided the phase condition (see the last equation in (4)) is replaced by  $\int_0^1 x^\top \dot{v} d\tau = 0$  at the first Newton correction. Otherwise,  $x$  would be undetermined among time-shifted solutions.

### Branch Switching at Flip Points

At a flip bifurcation point  $y^{\text{PD}} = (x^{\text{PD}}, p^{\text{PD}})$ , where  $x^{\text{PD}} \in C^1([0, 1], \mathbb{R}^n)$ ,  $x^{\text{PD}}(1) = x^{\text{PD}}(0)$ , one typically wants to

start the continuation of the emanating branch of “period-2” limit cycles, i.e., those which close to  $y^{\text{PD}}$  have approximately the double of the period of the bifurcating cycle. For this, branch switching at simple BPs can be used. In fact, two solution branches of the limit cycle BVP (4) transversely intersect at  $y^{\text{PD}}$  if one considers  $T$  as the doubled period: the branch of interest and the branch along which the corresponding period-1 cycle is traced twice. In other words, one can see the period-doubling bifurcation in the period-1 branch as the “period-halving” bifurcation in the period-2 branch.

Alternatively, the vector  $\phi = (v, q)$  tangent to the period-2 branch at  $y^{\text{PD}}$  is given by the flip theorem [56] and does not need to be computed by solving the algebraic branching Eq. (41). In particular, the initial solution of the period-2 BVP is  $x(t) = x^{\text{PD}}(2t)$ ,  $p = p^{\text{PD}}$ ,  $T = 2T^{\text{PD}}$ , while  $q = (p_s, T_s) = 0$  and

$$v(t) = \begin{cases} w(t), & 0 \leq t < 1, \\ -w(t-1), & 1 \leq t < 2, \end{cases}$$

where  $w(t)$  is the unit-length eigenfunction of the linearized (time-dependent) ODE associated with the multiplier  $-1$ , i.e.,

$$\begin{cases} \dot{w} - T^{\text{PD}} f_x(x^{\text{PD}}, p^{\text{PD}})w = 0, \\ w(1) + w(0) = 0, \\ \int_0^1 w(\tau)^\top w(\tau) d\tau = 1. \end{cases}$$

### Branch Switching at Codimension 2 Equilibria

Assume that  $y_2 = (x_2, p_2)$ ,  $x_2 \in \mathbb{R}^n$ ,  $p_2 \in \mathbb{R}^2$ , identifies a codim 2 equilibrium bifurcation point of system (1), either cusp (CP), Generalized Hopf (GH), Bogdanov-Takens, (BT), zero-Hopf (ZH), or double Hopf (HH). Then, according to the analysis of the corresponding normal form (13) with two parameters and critical dimension  $n_c = 1, 2, 3, 4$  at CP, BT and GH, ZH, HH points, respectively, several curves of codim 1 bifurcations of equilibria and limit cycles emanate from  $p_2$  in the parameter plane [56]. Here, the problem of how the continuation of such curves can be started from  $y_2$  is discussed, by restricting the attention to equilibria bifurcations (an example of a cycle bifurcation is given at the end, and see [9,61]).

In general, the normal form analysis also provides an approximation of each emanating codim 1 bifurcation, in the form of a parameterized expansion

$$w = \sum_{\mu \geq 1} \frac{1}{\mu!} w_\mu \varepsilon^\mu, \quad \alpha = \sum_{v \geq 1} \frac{1}{v!} \alpha_v \varepsilon^v, \quad (45)$$

for small  $\varepsilon > 0$  and up to some finite order. Then, the (parameter-dependent) center manifold (15) maps such an

approximation back to the solution branch of the original system (1), locally to  $y_2 = (H(0, 0), V(0))$ , and allows one to compute the proper prediction from  $y_2$  along the direction of the desired codim 1 bifurcation. However, as will be concluded in the following, such approximations are not needed to start equilibria bifurcations and are therefore not derived (see [9] for all available details).

**Switching at a Cusp Point** At a generic cusp point  $y^{\text{CP}} = (u^{\text{CP}}, p^{\text{CP}})$ , two fold bifurcation curves terminate tangentially. Generically, the cusp point is a regular point for the fold defining systems (24) and (25), where the tangent vector has zero  $p$ -components. The cusp geometrically appears once the fold solution branch is projected in the parameter plane. Thus, the continuation of the two fold bifurcations can simply be started as forward and backward fold continuation from  $y^{\text{CP}}$ . Since the continuation direction is uniquely defined, neither a tangent prediction nor a nonlinear expansion of the desired solution branch are necessary.

**Switching at a Generalized Hopf Point** At a Hopf point with vanishing Lyapunov coefficient  $y^{\text{GH}} = (x^{\text{GH}}, p^{\text{GH}})$ , a LPC bifurcation terminates tangentially to a Hopf bifurcation, which turns from super- to subcritical, and vice versa. Thus,  $y^{\text{GH}}$  is a regular point for the Hopf defining systems (27)–(30).

**Switching at a Bogdanov–Takens Point** At a generic Bogdanov–Takens point  $y^{\text{BT}} = (u^{\text{BT}}, p^{\text{BT}})$ , a Hopf and a (saddle) homoclinic bifurcation terminate tangentially to a fold bifurcation, along which a real nonzero eigenvalue of the Jacobi matrix  $f_x(x, p)$  changes sign at  $y^{\text{BT}}$ . Generically,  $y^{\text{BT}}$  is a regular point for the fold defining systems (24) and (25) and for the Hopf defining systems (27), (29), and (30). However,  $y^{\text{BT}}$  is a simple BP for the Hopf defining system (28). In fact, the fold and Hopf branches are both solution branches of the continuation problem (28), where  $\omega = 0$  and  $v_1 = v_2$ , with  $v_1^\top v_1 = v_2^\top v_2 = 1/2$ , along the fold branch. The branch switch procedures described in this section readily apply in this case.

**Switching at Zero-Hopf and Double Hopf Points** Generically, zero-Hopf and double Hopf points are regular points for codim 1 equilibria bifurcations (fold and Hopf), so that the proper initial prediction is uniquely defined. For limit cycle bifurcations and connecting orbits, nonlinear expansions of the fold and Hopf branches are needed to derive initial predictions for the emanating branches. For cycles the switching procedure can be set up using the

center manifold [61]. However, initial predictions for homoclinic and heteroclinic bifurcations for both zero-Hopf and double Hopf cases are not available in general, but see [35].

The double Hopf bifurcation appears when two different branches of Hopf bifurcations intersect. Several bifurcation curves are rooted at the double Hopf point. In particular, it is known that there are generically two branches, two half-lines in the parameter plane, emanating from this point along which a Neimark–Sacker bifurcation of limit cycles occurs [56]. Here it is discussed how to initialize the continuation of a Neimark–Sacker branch (using (37)) from a double Hopf point after continuation of a Hopf branch. The initialization requires approximations of the cycle  $x$ , the period  $T$ , the parameters  $p$  and the real part of the multiplier  $\hat{\kappa}$ . These can be obtained by reducing the dynamics of (1) to the center manifold. On the center manifold, the dynamics near a HH bifurcation point is governed by the following normal form

$$\begin{pmatrix} \dot{w}_1 \\ \dot{w}_2 \end{pmatrix} = \begin{pmatrix} (i\omega_1(\alpha) + \alpha_1)w_1 + g_{2100}w_1|w_1|^2 + g_{1011}w_1|w_2|^2 \\ (i\omega_2(\alpha) + \alpha_2)w_2 + g_{1110}w_2|w_1|^2 + g_{0021}w_2|w_2|^2 \end{pmatrix} + O(\| (w_1, w_2) \|^4), \quad (46)$$

where  $(w_1, w_2) \in \mathbb{C}^2$ . In polar coordinates,  $w_1 = \rho_1 e^{i\theta_1}$ ,  $w_2 = \rho_2 e^{i\theta_2}$ , the asymptotics from the normal form as in (45) for the nonhyperbolic cycle on one branch are given by

$$(\rho_1, \rho_2, \alpha_1, \alpha_2) = (\varepsilon, 0, -\Re e(g_{2100})\varepsilon^2, -\Re(g_{1110})\varepsilon^2), \quad (47)$$

with  $\theta_1 \in [0, 2\pi]$ ,  $\theta_2 = 0$ .

Although high-dimensional, the computation of the coefficients and center manifold vectors is relatively straightforward. Introduce  $Av_j = i\omega_j v_j$ ,  $A^\top w_j = -i\omega_j w_j$  with  $\bar{v}_j^\top v_j = \bar{w}_j^\top w_j = 1$  and let  $v = (10), (01)$  and introduce the standard basis vectors  $e_{10} = (1, 0)$ ,  $e_{01} = (0, 1)$ . Using the expansion (14), the cubic critical normal form coefficients and the parameter dependence are calculated from

$$\begin{aligned} g_{2100} &= \bar{w}_1^\top [C(v_1, v_1, \bar{v}_1) + B(h_{2000}, \bar{v}_1) \\ &\quad + 2B(h_{1100}, v_1)]/2, \\ g_{1011} &= \bar{w}_1^\top [C(v_1, v_2, \bar{v}_2) + B(h_{1010}, \bar{v}_2) \\ &\quad + B(h_{1001}, v_2) + B(h_{0011}, v_1)], \end{aligned}$$

$$\begin{aligned}
g_{1110} &= \bar{w}_2^\top [C(v_2, v_1, \bar{v}_1) + B(h_{1100}, v_2) + B(h_{1010}, \bar{v}_1) \\
&\quad + B(\bar{h}_{1001}, v_1)] , \\
g_{0021} &= \bar{w}_2^\top [C(v_2, v_2, \bar{v}_2) + B(h_{0020}, \bar{v}_2) \\
&\quad + 2B(h_{0011}, v_2)]/2 , \\
\Gamma_{j,v} &= \bar{p}_j^\top [A_1(v_j, e_v) - B(v_j, A^{-1}J_1 e_v)]/2 , \quad (48)
\end{aligned}$$

where  $j = 1, 2$  and

$$\begin{aligned}
h_{2000} &= (2i\omega_1 I_n - A)^{-1} B(v_1, v_1) , \\
h_{0020} &= (2i\omega_2 I_n - A)^{-1} B(v_2, v_2) , \\
h_{1100} &= -A^{-1} B(v_1, \bar{v}_1) , \\
h_{0011} &= -A^{-1} B(v_2, \bar{v}_2) , \\
h_{1010} &= (i(\omega_1 + \omega_2) I_n - A)^{-1} B(v_1, v_2) , \\
h_{1001} &= (i(\omega_1 - \omega_2) I_n - A)^{-1} B(v_1, \bar{v}_2) , \quad (49)
\end{aligned}$$

where  $h_\mu$  are the vectors in the expansion of the center manifold (17).

Now, to construct a cycle for (46), a mesh for  $\theta_1$  is defined and the asymptotics (47) are inserted into the polar coordinates. This cycle is mapped back to the original system by using (17) with (47), (48) and (49). The transformation between free system and unfolding parameters is given by  $p(\varepsilon) = \mathcal{R}(\Gamma_1 \Gamma_2)^{-1}(\alpha_1, \alpha_2)^\top$  using (47). Finally, approximating formulas for the period and the real part of the multiplier are given by

$$\begin{aligned}
T &= \frac{2\pi}{\omega_1 + d\omega_1 \varepsilon^2} , \quad \hat{\kappa} = \cos(T(\omega_2 + d\omega_2 \varepsilon^2)) , \\
(d\omega_1, d\omega_2) &= -\Im(\Gamma_1 \Gamma_2)^\top (\Re(\Gamma_1 \Gamma_2)^\top)^{-1} \Re(g_{2100}, g_{1110})^\top \\
&\quad + \Im(g_{2100}, g_{1110}) ,
\end{aligned}$$

where  $d\omega_1, d\omega_2$  indicate the change in rotation in the angles  $\theta_1, \theta_2$  for  $\varepsilon \neq 0$ . This construction is done up to second order in  $\varepsilon$  and leads to an initial approximation for the continuation of a Neimark–Sacker bifurcation curve starting from a double Hopf point. A similar set up can be defined for the other branch.

## Connecting Orbits

Connecting orbits, such as homoclinic and heteroclinic orbits, can be continued using a variety of techniques. To fix some terminology: a heteroclinic orbit that connects two equilibrium points  $x_-$  and  $x_+$  in the ODE system (1) is an orbit for which

$$\lim_{t \rightarrow -\infty} x(t) = x_- \quad \text{and} \quad \lim_{t \rightarrow \infty} x(t) = x_+ .$$

A homoclinic orbit is an orbit connecting an equilibrium point to itself, that is, if  $x_+ = x_-$ . Similarly there exist heteroclinic connecting orbits between equilibrium points and periodic orbits, and homoclinic and heteroclinic orbits connecting periodic orbits to periodic orbits.

Traditionally, homoclinic orbits to equilibrium points were computed indirectly using numerical shooting or by continuing a periodic solution with a large enough but fixed period, that is close enough to a homoclinic orbit [24]. More modern and robust techniques compute connecting orbits directly using *projection boundary conditions*. Computing connections with and between periodic orbits is subject to current research [26,27,54]. For a more detailed description of the methods described here, see [9].

## Formulation as a BVP

A heteroclinic orbit can be expressed as a BVP in the following way:

$$\begin{aligned}
\dot{x}(t) &= f(x(t), p) , \quad \lim_{t \rightarrow -\infty} x(t) = x_- , \\
\lim_{t \rightarrow \infty} x(t) &= x_+ , \quad \int_{-\infty}^{\infty} (x(t) - x_0(t))^\top \dot{x}_0(t) dt = 0 ,
\end{aligned}$$

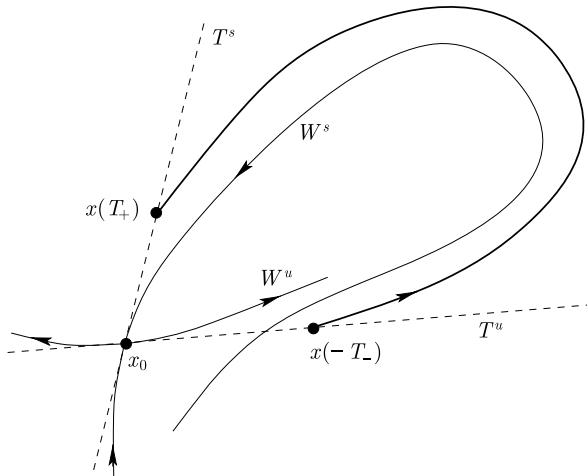
where the integral condition is with respect to a reference solution  $x_0(t)$  and fixes the phase, similarly to the phase condition for periodic orbits. This BVP, however, operates on an infinite interval, while, numerically, one can only operate on a finite, truncated interval  $[-T_-, T_+]$ . In this case the problem can be reformulated as

$$\begin{cases} \dot{x}(t) - f(x(t), p) = 0 , \\ L_s(p)(x(-T_-) - x_-(p)) = 0 , \\ L_u(p)(x(T_+) - x_+(p)) = 0 , \\ \int_{-T_-}^{T_+} (x(t) - x_0(t))^\top \dot{x}_0(t) dt = 0 , \end{cases}$$

where the equations involving  $L_s(p)$  and  $L_u(p)$  form the projection boundary conditions. Here  $L_s(p)$  is an  $n_s \times n$  matrix where the rows span the  $n_s$ -dimensional stable eigenspace of  $A^\top(x_-)$ , and similarly  $L_u(p)$  is an  $n_u \times n$  matrix where the rows span the  $n_u$ -dimensional unstable eigenspace of  $A^\top(x_+)$ , where  $A(x)$  denotes the Jacobi matrix of (1) at  $x$ . The projection boundary conditions then ensure that the starting point  $x(-T_-)$  lies in the unstable eigenspace of  $x_-$  and that the end point  $x(T_+)$  lies in the stable eigenspace of  $x_+$  (see Fig. 5).

In parallel, one must also continue of the equilibrium points  $x_-(p)$  and  $x_+(p)$ , unless they are fixed. Then the eigenspaces can be determined by doing a Schur decomposition of the corresponding Jacobi matrices. These must





**Numerical Bifurcation Analysis, Figure 5**

Projection boundary conditions in two dimensions: the orbit homoclinic to  $x_0$  is approximated by the truncated orbit from  $x(-T_-)$  on the unstable eigenspace  $T^u$  to  $x(T_+)$  on the stable eigenspace  $T^s$ . The unstable and stable manifolds are denoted by  $W^u$  and  $W^s$ , respectively: this figure shows that the homoclinic orbit is also approximated in parameter space, because the two manifolds do not coincide

be subsequently scaled to ensure continuity in the parameter  $p$  [6]. Alternatively, one can construct smooth projectors using an algorithm for the continuation of invariant subspaces which includes the Riccati equation in the defining system, see [33] for this new method. All these conditions taken together give a BVP with two free parameters, since, in general, a homoclinic or heteroclinic connection is a codim 1 phenomenon.

### Detecting Homoclinic Bifurcations

It is then possible to detect codim 2 bifurcations of homoclinic orbits by setting up test functions and monitoring those, detecting when they cross zero. For the continuation of these codim 2 bifurcations in three parameters such test functions can then be kept constant equal to zero, which provides an extra boundary condition. Simple test functions involve the values of leading eigenvalues of the equilibrium point, i.e., the stable and unstable eigenvalues closest to the imaginary axis. Some other bifurcations such as the inclination flip involve solving the so-called *adjoint variational equation*, which can detect whether the flow around the orbit is orientable or twisted like a Möbius strip. Homoclinics to a saddle-node, that is, where one of the eigenvalues of the equilibrium is zero, can be detected and followed similarly, by constructing appropriate test functions. For details, see [9] and the references mentioned therein.

### Homoclinic Branch Switching

It is sometimes desirable to do branch switching from a homoclinic orbit to an  $n$ -homoclinic orbit, that is, a homoclinic orbit that goes through some neighborhood of the related equilibrium point  $n - 1$  times before finally converging to the equilibrium. Such  $n$ -homoclinic orbits arise in a number of situations involving the eigenvalues of the equilibrium and the orientation of the flow around the orbit. Suppose that the orbit is homoclinic to a saddle with complex conjugate eigenvalues (a saddle-focus). Let the *saddle-quantity*  $\sigma$  be the sum of the real parts of the leading stable and the leading unstable eigenvalue. If this quantity is positive, then a so-called Shil'nikov snake exists, which implies the existence of infinitely many  $n$ -periodic and  $n$ -homoclinic orbits for nearby parameters. These  $n$ -homoclinic orbits also arise from certain codim 2 bifurcations:

1. Belyakov bifurcations: Either the saddle-quantity of the saddle-focus goes through zero, or there is a transition between a saddle-focus and a real saddle (here the equilibrium has a double leading eigenvalue).
2. Homoclinic flip bifurcations: The inclination flip and the orbit flip, where the flow around the orbit changes between orientable and twisted.
3. The resonant homoclinic doubling: A homoclinic orbit for which the flow is twisted connected to a *real* saddle where the saddle quantity  $\sigma$  goes through zero.

Case 1. produces infinitely many  $n$ -homoclinic orbits, whereas cases 2. and 3. only produce a two-homoclinic orbit.

By breaking up a homoclinic orbit globally into two parts where the division is in a cross-section away from the equilibrium, somewhere “half-way”, and then gluing pieces together, it is possible to construct  $n$ -homoclinic orbits from 1-homoclinic orbits. The gaps between to-be-glued pieces can be well-defined using Lin’s method, which leaves the gaps in the direction of the adjoint vector. This vector can be found by solving the adjoint variational equation mentioned above at the “half-way” point. Combining gap distances and times taken for the flow in pieces, one can construct a well-posed BVP. This BVP can then be continued in those quantities so that if the gap sizes go to zero, one converges to an  $n$ -homoclinic orbit [66].

Lin’s method was also applied in [54] to compute point-to-cycle connections by gluing a piece from the equilibrium to a cross-section to a piece from the cross section to the cycle.

## Software Environments

This review has outlined necessary steps and suitable methods to perform numerical bifurcation analysis. Summarizing, the following subsequent tasks can be recognized.

Initial procedure	1. Compute the invariant solution
Continuation	2. Characterize the linearized behavior
	3. Variation of parameters
	4. Monitor dynamical indicators
Automated analysis	5. Detect special points and compute normal forms
	6. Switch branches

Ideally, these computations are automatically performed by software and indeed, many efforts have been spent on implementing the algorithms mentioned in this review and related ones. With the appearance of computers at research institutes the first codes and noninteractive packages were developed. For a recent historical overview of these and their capabilities and algorithms, see [38]. Here it is worthwhile to mention AUTO86 [24], LINBLF [51] and DSTOOL [4] as these are the predecessors to the packages AUTO-07P [21], CONTENT [58], MATCONT [19], PyDSTOOL [13] discussed here. In particular, AUTO86 is very powerful and in use to date, but not always easy to handle. Therefore, several attempts have been made to make an interactive graphical user interface. One example is XPPAUT [32] which is a standard tool in neuroscience.

The latest version of AUTO is AUTO-07P and written in Fortran and supports parallelization. It uses pseudo-arclength continuation and performs a limited bifurcation analysis of equilibria, fixed points, limit cycles and connecting orbits using HOMCONT [12]. A recent addition is the continuation of branch points [16]. It uses fully extended systems, but has specially adapted linear solvers so that it is still quite fast.

An interactive package written in C++ is CONTENT, where the Moore–Penrose continuation was first implemented. It supports bifurcation analysis of equilibria up to the continuation of codim 2 and detection of some codim 3 bifurcations. It uses a similar procedure as AUTO to continue limit cycles and detects codim 1 bifurcations of limit cycles, but does not continue these. It also handles bifurcations of cycles of maps up to the detection of codim 2 bifurcations. Normal forms for all codim 1 bifurcations are computed. Interestingly, both fully and minimally extended systems are implemented.

A new project MATCONT emerged out of CONTENT. It is written in MATLAB, a widely used software tool in mod-

eling. In contrast to AUTO it uses Moore–Penrose continuation and minimally extended systems to compute equilibria, cycles and connecting orbits. Since MATLAB is an interpreted language, it is slower than the other packages, although a considerable speedup is obtained as the code for the Jacobi matrices for limit cycles and their bifurcations are written in C-codes and compiled. MATCONT has, however, much more functionality. It computes normal form coefficients up to codim 2 bifurcations of equilibria [20] and for codim 1 of limit cycles normal form coefficients are computed using a BVP-algorithm [28]. A recent addition is to switch branches from codim 2 bifurcation points of equilibria to codim 1 bifurcations of cycles [61].

Finally PyDSTOOL supports similar functionality for ODEs as CONTENT.

## Future Directions

This review focuses on methods for equilibria and limit cycles to gain insight into the dynamics of a nonlinear ODE depending on parameters, but, generally speaking, other characteristics play a role too. For instance, a Neimark–Sacker bifurcation leads to the presence of tori. The computation and continuation of higher dimensional tori has been considered in [49,70,71]. The generalization of the methods for equilibria and cycles is, however, not straightforward. Stable and unstable manifolds are another aspect. In particular their visualization can hint at the presence of global bifurcations. A review of methods for computing such manifolds is presented in [52]. Connecting orbits can be calculated, but initializing such a continuation is a non-trivial task. This may be started from certain codim 2 bifurcation points as in [8], but good initial approximations are not available for other cases.

This review is also restricted to ODEs, but one can define other classes of dynamical systems. For instance, if the system is given by an explicitly defined map, i. e., not implicitly as a Poincaré map, the described approach can also be carried out [10,37,40,59]. Another important and related class is given by delay equations, and the algorithms for equilibria and periodic orbits in the ODE case can be applied with suitable modifications, see [5,74] and the software implementations DDE-BIFTOOL [31] and PDDE-CONT. The computation of normal forms and connecting orbits for this class is not yet thoroughly investigated or supported.

For large systems, e. g., discretizations of a partial differential equation (PDE), the algebra for some algorithms in this review becomes quite involved and numerically expensive. These problems need special treatment,

see LOCA [69], PDE-CONT [30,62] and related references. These packages focus on computing (periodic) solutions and bifurcation curves. Good algorithms for analyzing bifurcations of PDEs will be a major research topic.

Finally, there are also slow-fast (stiff) systems or systems with a particular structure such as symmetries. For these classes, many questions remain open.

While on most of the mentioned topics pioneering work has been done, the methods are far from as “complete” as for ODEs. One overview of current research topics in dynamical systems can be found in [53].

## Bibliography

### Primary Literature

- Allgower EL, Georg K (2000) *Numerical Continuation Methods: An Introduction*. Springer, Berlin
- Arnold VI (1983) *Geometrical Methods in the Theory of Ordinary Differential Equations*. Springer, Berlin
- Ascher UC, Mattheij RMM, Russell B (1995) *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. SIAM, Philadelphia
- Back A, Guckenheimer J, Myers M, Wicklin F, Worfolk P (1992) DsTool: Computer assisted exploration of dynamical systems. *Notices Amer Math Soc* 39:303–309
- Barton DAW, Krauskopf B, Wilson RE (2007) Homoclinic bifurcations in a neutral delay model of a transmission line oscillator. *Nonlinearity* 20:809–829
- Beyn WJ (1990) The numerical computation of connecting orbits in dynamical systems. *IMA J Numer Anal* 10:379–405
- Beyn WJ (1991) Numerical methods for dynamical systems. In: Light W (ed) *Advances in numerical analysis, Vol. I* (Lancaster, 1990). Clarendon, Oxford, pp 175–236
- Beyn WJ (1994) Numerical analysis of homoclinic orbits emanating from a Takens–Bogdanov point. *IMA J Numer Anal* 14:381–410
- Beyn WJ, Champneys A, Doedel EJ, Govaerts W, Kuznetsov YA, Sandstede B (2002) Numerical continuation, and computation of normal forms. In: Fiedler B (ed) *Handbook of Dynamical Systems*, vol 2. Elsevier Science, Amsterdam, pp 149–219
- Beyn WJ, Kleinkauf JM (1997) The numerical computation of homoclinic orbits for maps. *Siam J Numer Anal* 34:1207–1236
- Carr J (1981) *Applications of centre manifold theory*. Springer, New York
- Champneys AR, Kuznetsov YA, Sandstede B (1996) A numerical toolbox for homoclinic bifurcation analysis. *Int J Bif Chaos* 6:867–887
- Clewley R, Sherwood E, LaMar D, Guckenheimer J (2005) PyDSTool, available at <http://sourceforge.net/projects/pydstool>. Accessed 09 Sept 2008
- Coullet PH, Spiegel EA (1983) Amplitude equations for systems with competing instabilities. *SIAM J Appl Math* 43:776–821
- de Boor C, Swartz B (1973) Collocation at Gaussian points. *SIAM J Num Anal* 10:582–606
- Dercole F (2008) Bpcont: An auto driver for the continuation of branch points of algebraic and boundary-value problems. *SIAM J Sci Comp* 30:2405–2426
- Deuflhard P, Bornemann F (2002) *Scientific Computing With Ordinary Differential Equations*. In: Marsden JE, Sirovich L, Antman S (eds) *Texts in applied mathematics*, vol 42. Springer, New York
- Deuflhard P, Fiedler B, Kunkel P (1987) Efficient numerical pathfollowing beyond critical points. *SIAM J Num Anal* 24:912–927
- Dhooge A, Govaerts W, Kuznetsov YA (2003) Matcont: A Matlab package for numerical bifurcation analysis of ODE's. *ACM TOMS* 29, pp 141–164. <http://sourceforge.net/projects/matcont>. Accessed 9 Sep 2008
- Dhooge A, Govaerts W, Kuznetsov YA, Meijer HGE, Sautois B (2008) New features of the software Matcont for bifurcation analysis of dynamical systems. *Math Comp Mod Dyn Syst* 14:147–175
- Doedel EJ (2007) AUTO-07P: Continuation and Bifurcation Software for Ordinary Differential Equations, User's Guide. <http://cmvl.cs.concordia.ca/auto>. Accessed 9 Sep 2008
- Doedel EJ (2007) Lecture notes on numerical analysis of nonlinear equations. In: Krauskopf B, Osinga HM, Galan-Vioque J (eds) *Numerical continuation methods for dynamical systems: Path following and boundary value problems*. Springer-Canopus, Dordrecht, pp 1–49
- Doedel EJ, Govaerts W, Kuznetsov YA (2003) Computation of periodic solution bifurcations in ODEs using bordered systems. *SIAM J Numer Anal* 41:401–435
- Doedel EJ, Kernevez JP (1986) AUTO, Software for Continuation and Bifurcation Problems in Ordinary Differential Equations. *Applied Mathematics*, California Institute of Technology, Pasadena
- Doedel EJ, Romanov VA, Paffenroth RC, Keller HB, Dichmann DJ, Vioque GJ, Vanderbauwhede A (2007) Elemental periodic orbits associated with the liberation points in the circular restricted 3-body problem. *Int J Bif Chaos* 17:2625–2677
- Doedel EJ, Kooi BW, Kuznetsov YA, van Voorn GAK (2008) Continuation of connecting orbits in 3D-ODEs: (I) Point-to-cycle connections. *Int J Bif Chaos* 18:1889–1903
- Doedel EJ, Kooi BW, Kuznetsov YA, van Voorn GAK (2008) Continuation of connecting orbits in 3D-ODEs: (II) Cycle-to-cycle connections. [arXiv.org/0804.0179](http://arXiv.org/0804.0179). Accessed 9 Sep 2008
- Doedel EJ, Kuznetsov YA, Govaerts W, Dhooge A (2005) Numerical continuation of branch points of equilibria and periodic orbits. *Int J Bif Chaos* 15:841–860
- Elphick C, Tirapegui E, Brachet M, Coulet P, looss G (1987) A simple global characterization for normal forms of singular vector fields. *Phys D* 32:95–127
- Engelborghs K, Lust K, Roose D (1999) Direct computation of period doubling bifurcation points of large-scale systems of ODEs using a Newton-Picard method. *IMA J Numer Anal* 19:525–547
- Engelborghs K, Luzyanina T, Roose D (2002) Numerical bifurcation analysis of delay differential equations using DDE-BIFTOOL. *ACM Trans Math Softw* 28:1–21
- Ermentrout B (2002) *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*. SIAM, Philadelphia
- Friedman M, Govaerts W, Kuznetsov YA, Sautois B (2005) Continuation of homoclinic orbits in MATLAB. In: Sunderam VS, van Albada GD, Sloot PMA, Dongarra J (eds) *Proceedings ICCS 2005, Atlanta, Part I*, Springer Lecture Notes in Computer Science Vol. 3514. Springer, Berlin, pp 263–270

34. Fuller AT (1968) Conditions for a matrix to have only characteristic roots with negative real parts. *J Math Anal Appl* 23: 71–98
35. Gaspard P (1993) Local birth of homoclinic chaos. *Phys D* 62:94–122
36. Govaerts W, Guckenheimer J, Khibnik A (1997) Defining functions for multiple Hopf bifurcations. *SIAM J Numer Anal* 34:1269–1288
37. Govaerts W, Khoshshar Ghaziani R, Kuznetsov YA, Meijer HGE (2007) Numerical methods for two-parameter local bifurcation analysis of maps. *SIAM J Sci Comput* 29:2644–2667
38. Govaerts W, Kuznetsov YA (2007) Numerical continuation tools. In: Krauskopf B, Osinga HM, Galan-Vioque J (eds) *Numerical continuation methods for dynamical systems: Path following and boundary value problems*. Springer-Canopus, Dordrecht, pp 51–75
39. Govaerts W, Kuznetsov YA, Dhooge A (2005) Numerical continuation of bifurcations of limit cycles in matlab. *SIAM J Sci Comp* 27:231–252
40. Govaerts W, Kuznetsov YA, Sijnave B (1999) Bifurcations of maps in the software package content. In: Ganzha VG, Mayr EW, Vorozhtsov EV (eds) *Computer algebra in scientific computing—CASC’99* (Munich). Springer, Berlin, pp 191–206
41. Govaerts W, Pryce JD (1993) Mixed block elimination for linear systems with wider borders. *IMA J Num Anal* 13:161–180
42. Govaerts WJF (2000) *Numerical Methods for Bifurcations of Dynamical Equilibria*. SIAM, Philadelphia
43. Guckenheimer J (2002) Numerical analysis of dynamical systems. In: Fiedler B (ed) *Handbook of dynamical systems*, Vol. 2. Elsevier Science, North-Holland, pp 346–390
44. Guckenheimer J, Holmes P (1983) *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer, New York
45. Guckenheimer J, Myers M, Sturmfels B (1997) Computing Hopf bifurcations I. *SIAM J Numer Anal* 34:1–21
46. Henderson ME (2007) Higher-dimensional continuation. In: Krauskopf B, Osinga HM, Galan-Vioque J (eds) *Numerical continuation methods for dynamical systems: Path following and boundary value problems*. Springer-Canopus, Dordrecht, pp 77–115
47. Iooss G (1988) Global characterization of the normal form for a vector field near a closed orbit. *J Diff Eqs* 76:47–76
48. Iooss G, Adelmeyer M (1992) *Topics in Bifurcation Theory and Applications*. World Scientific, Singapore
49. Jorba A (2001) Numerical computation of the normal behaviour of invariant curves of  $n$ -dimensional maps. *Nonlinearity* 14:943–976
50. Keller HB (1977) Numerical solution of bifurcation and nonlinear eigenvalue problems. In: Rabinowitz PH (ed) *Applications of bifurcation theory*. Proc Advanced Sem, Univ Wisconsin, Madison, (1976), Publ Math Res Center, No. 38. Academic Press, New York pp 359–384
51. Khibnik AI (1990) LINBLF: A program for continuation and bifurcation analysis of equilibria up to codimension three. In: Roose D, de Dier D, Spence A (eds) *Continuation and bifurcations: numerical techniques and applications*. Leuven, (1989), NATO Adv Sci Inst Ser C Math Phys Sci, vol 313. Kluwer, Dordrecht, pp 283–296
52. Krauskopf B, Osinga HM, Doedel EJ, Henderson ME, Guckenheimer J, Vladimirovsky A, Dellnitz M, Junge O (2005) A survey of methods for computing (un)stable manifolds of vector fields. *Int J Bif Chaos* 15(3):763–791
53. Krauskopf B, Osinga HM, Galan-Vioque J (2007) *Numerical Continuation methods for dynamical systems*. Springer-Canopus, Dordrecht
54. Krauskopf B, Riess T (2008) A Lin’s method approach to finding and continuing heteroclinic connections involving periodic orbits. *Nonlinearity* 21:1655–1690
55. Kuznetsov YA (1999) Numerical normalization techniques for all codim 2 bifurcations of equilibria in ODE’s. *SIAM J Numer Anal* 36:1104–1124
56. Kuznetsov YA (2004) *Elements of Applied Bifurcation Theory*, 3rd edn. Springer, Berlin
57. Kuznetsov YA, Govaerts W, Doedel EJ, Dhooge (2005) Numerical periodic normalization for codim 1 bifurcations of limit cycles. *SIAM J Numer Anal* 43:1407–1435
58. Kuznetsov YA, Levitin VV (1995) Content: A multiplatform environment for analyzing dynamical systems. <http://ftp.cwi.nl/pub/CONTENT/>. Accessed 9 Sep 2008
59. Kuznetsov YA, Meijer HGE (2005) Numerical normal forms for codim 2 bifurcations of maps with at most two critical eigenvalues. *SIAM J Sci Comput* 26:1932–1954
60. Kuznetsov YA, Meijer HGE (2006) Remarks on interacting Neimark–Sacker bifurcations. *J Diff Eqs Appl* 12:1009–1035
61. Kuznetsov YA, Meijer HGE, Govaerts W, Sautois B (2008) Switching to nonhyperbolic cycles from codim 2 bifurcations of equilibria in ODEs. *Physica D* 237:3061–3068
62. Lust K, Roose D, Spence A, Champneys AR (1998) An adaptive Newton-Picard algorithm with subspace iteration for computing periodic solutions. *SIAM J Sci Comput* 19:1188–1209
63. Mei Z (1989) A numerical approximation for the simple bifurcation problems. *Numer Func Anal Opt* 10:383–400
64. Mei Z (2000) *Numerical Bifurcation Analysis for Reaction-Diffusion Equations*. Springer, Berlin
65. Moore G (1980) The numerical treatment of nontrivial bifurcation points. *Numer Func Anal Opt* 2:441–472
66. Oldeman BE, Champneys AR, Krauskopf B (2003) Homoclinic branch switching: a numerical implementation of Lin’s method. *Int J Bif Chaos* 13:2977–2999
67. Roose D, Hlaváček V (1985) A direct method for the computation of Hopf bifurcation points. *SIAM J Appl Math* 45:879–894
68. Russell RD, Christiansen J (1978) Adaptive mesh selection strategies for solving boundary value problems. *SIAM J Numer Anal* 15:59–80
69. Salinger AG, Burroughs EA, Pawlowski RP, Phipps ET, Romero LA (2005) Bifurcation tracking algorithms and software for large scale applications. *Int J Bif Chaos* 15:1015–1032
70. Schilder F, Osinga HM, Vogt W (2005) Continuation of quasiperiodic invariant tori. *SIAM J Appl Dyn Syst* 4:459–488
71. Schilder F, Vogt W, Schreiber S, Osinga HM (2006) Fourier methods for quasi-periodic oscillations. *Int J Num Meth Eng* 67:629–671
72. Shoshitaishvili AN (1975) The bifurcation of the topological type of the singular points of vector fields that depend on parameters. *Trudy Sem Petrovsk*, (Vyp. 1):279–309
73. Stephanos C (1900) Sur une extension du calcul des substitutions linéaires. *J Math Pures Appl* 6:73–128
74. Szalai R, Stépán G, Hogan SJ (2006) Continuation of bifurcations in periodic delay-differential equations using characteristic matrices. *SIAM J Sci Comput* 28:1301–1317

75. van Strien SJ (1979) Center manifolds are not  $C^\infty$ . *Math Zeitschrift* 166:143–145

### Books and Reviews

- Devaney RL (2003) *An introduction to chaotic dynamical systems*. Westview Press, Boulder
- Doedel EJ, Keller HB, Kernevez JP (1991) *Numerical Analysis and Control of Bifurcation Problems (I): Bifurcation in Finite Dimensions*. *Int J Bif Chaos* 1:493–520
- Doedel EJ, Keller HB, Kernevez JP (1991) *Numerical Analysis and Control of Bifurcation Problems (II): Bifurcation in Infinite Dimensions*. *Int J Bif Chaos* 1:745–772
- Hirsch MW, Smale S, Devaney RL (2004) *Differential equations, dynamical systems, and an introduction to chaos*, 2nd edn. *Pure and Applied Mathematics*, vol 60 Elsevier/Academic Press, Amsterdam
- Murdock J (2003) *Normal Forms and Unfoldings for Local Dynamical Systems*. Springer, New York

## Numerical Issues When Using Wavelets

JEAN-LUC STARCK<sup>1</sup>, JALAL FADILI<sup>2</sup>

<sup>1</sup> DAPNIA/SEDI-SAP, Service d'Astrophysique,  
CEA/Saclay, Gif sur Yvette, France

<sup>2</sup> GREYC CNRS UMR 6072, Image Processing Group,  
Ecole Nationale Supérieure d'Ingénieurs de Caen,  
Caen Cedex, France

### Article Outline

[Notations](#)

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Continuous Wavelet Transform](#)

[The \(Bi-\)Orthogonal Wavelet Transform](#)

[The Lifting Scheme](#)

[The Undecimated Wavelet Transform](#)

[The 2D Isotropic Undecimated Wavelet Transform](#)

[Designing Non-orthogonal Filter Banks](#)

[Iterative Reconstruction](#)

[Future Directions](#)

[Bibliography](#)

### Notations

For a real discrete-time filter whose impulse response is  $h[n]$ ,  $\hat{h}[n] = h[-n]$ ,  $n \in \mathbb{Z}$  is its time-reversed version. The  $\hat{\phantom{x}}$  notation will be used for the Fourier transform of square-integrable signals. For a filter  $h$ , its  $z$ -transform is written  $H(z)$ . The convolution prod-

uct of two signals in  $\ell^2(\mathbb{Z})$  will be written  $*$ . For the octave band wavelet representation, analysis (respectively, synthesis) filters are denoted  $h$  and  $g$  (respectively,  $\hat{h}$  and  $\hat{g}$ ). The scaling and wavelet functions used for the analysis (respectively, synthesis) are denoted  $\phi(\phi(x/2) = \sum_k h[k]\phi(x-k)$ ,  $x \in \mathbb{R}$  and  $k \in \mathbb{Z}$ ) and  $\psi(\psi(x/2) = \sum_k g[k]\phi(x-k)$ ,  $x \in \mathbb{R}$  and  $k \in \mathbb{Z}$ ) (respectively,  $\tilde{\phi}$  and  $\tilde{\psi}$ ). We also define the scaled dilated and translated version of  $\phi$  at scale  $j$  and position  $k$  as  $\phi_{j,k}(x) = 2^{-j}\phi(2^{-j}x - k)$ , and similarly for  $\psi$ ,  $\tilde{\phi}$  and  $\tilde{\psi}$ .

### Glossary

**WT** Wavelet Transform

**CWT** Continuous Wavelet Transform

**DWT** Discrete (decimated) Wavelet Transform

**UWT** Undecimated Wavelet Transform

**IUWT** Isotropic Undecimated Wavelet Transform

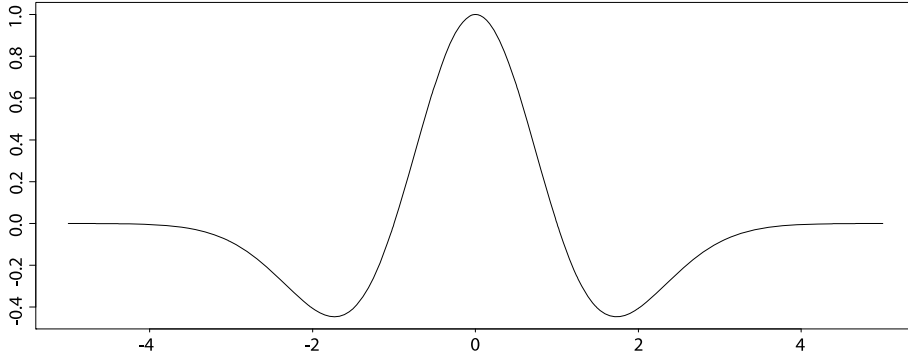
### Definition of the Subject

Wavelets and related multiscale representations pervade all areas of signal processing. The recent inclusion of wavelet algorithms in JPEG 2000 – the new still-picture compression standard – testifies to this lasting and significant impact. The reason of the success of the wavelets is due to the fact that wavelet basis represents well a large class of signals, and therefore allows us to detect roughly isotropic elements occurring at all spatial scales and locations. As the noise in the physical sciences is often not Gaussian, the modeling, in the wavelet space, of many kind of noise (Poisson noise, combination of Gaussian and Poisson noise, long-memory  $1/f$  noise, non-stationary noise, ...) has also been a key step for the use of wavelets in scientific, medical, or industrial applications [41]. Extensive wavelet packages exist now, commercial (see for example [29]) or non commercial (see for example [47,48]), which allows any researcher, doctor, or engineer to analyze his data using wavelets.

### Introduction

Over the last two decades there has been abundant interest in wavelet methods. In many hundreds of papers published in journals throughout the scientific and engineering disciplines, a wide range of wavelet-based tools and ideas have been proposed and studied. Background texts on the wavelet transform include [11,13,26,41,44]. The most widely used wavelet transform (WT) algorithm is certainly the decimated bi-orthogonal wavelet transform (DWT) which is used in JPEG 2000. While the bi-orthogonal wavelet transform has led to successful implementa-





Numerical Issues When Using Wavelets, Figure 1  
Mexican hat function

tion in image compression, results were far from optimal for other applications such as filtering, deconvolution, detection, or more generally, analysis of data. This is mainly due to the loss of the translation-invariance property in the DWT, leading to a large number of artifacts when an image is reconstructed after modification of its wavelet coefficients. Later efforts found that substantial improvements in perceptual quality could be obtained by translation invariant methods based on thresholding of an undecimated wavelet transform.

### The Continuous Wavelet Transform

The continuous wavelet transform uses a single function  $\psi(x)$  and all its dilated and shifted version to analyze signals. The Morlet–Grossmann definition [20] of the continuous wavelet transform (CWT) for a 1-dimensional real-valued function<sup>1</sup>  $f(x) \in L^2(\mathbb{R})$ , the space of all square-integrable functions, is:

$$W(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x) \psi\left(\frac{x-b}{a}\right) dx \quad (1)$$

where:

- $W(a, b)$  is the wavelet coefficient of the function  $f(x)$ ,
- $\psi(x)$  is the analyzing wavelet,
- $a (> 0)$  is the scale parameter,
- $b$  is the position parameter.

The inverse transform is obtained by:

$$f(x) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{a}} W(a, b) \times \psi\left(\frac{x-b}{a}\right) \frac{da db}{a^2} \quad (2)$$

<sup>1</sup>We only consider here real wavelets. This can be extended to complex wavelets without too much difficulty.

where:

$$C_\psi = \int_0^{+\infty} \frac{|\hat{\psi}|^2}{\nu} d\nu = \int_{-\infty}^0 \frac{|\hat{\psi}|^2}{\nu} d\nu. \quad (3)$$

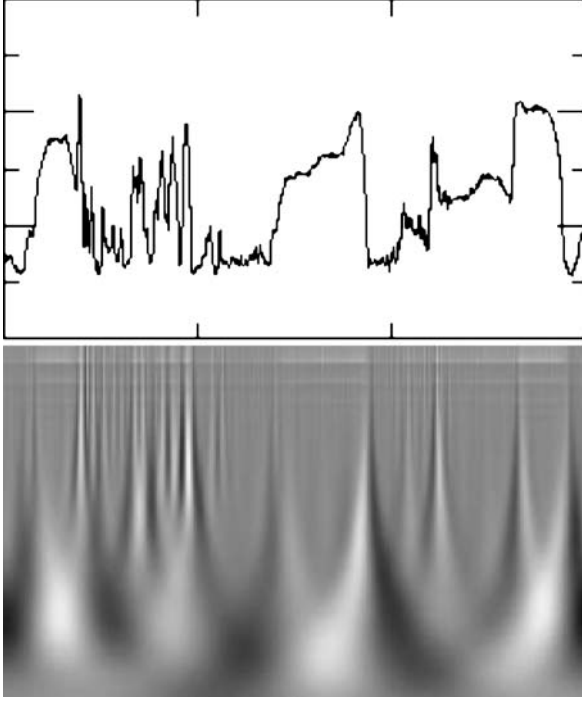
Reconstruction is only possible if  $C_\psi$  is finite (admissibility condition) which implies that  $\hat{\psi}(0) = 0$ , i. e. the mean of the wavelet function is 0. The wavelet is said to have a zero moment property, and appears to have a band-pass profile. A closely related relation to the inverse given in Eq. (2), is an energy conservation formula, an analogue to Plancherel's formula [26].

Figure 1 shows the Mexican hat wavelet function, which is defined by:

$$\psi(x) = (1 - x^2) e^{-x^2/2}. \quad (4)$$

This is the second derivative of a Gaussian. The lower-part of Fig. 2 shows the CWT of a 1D signal (top plot of Fig. 2) computed with the Mexican Hat wavelet. This diagram is called a *scalogram*. Its  $y$ -axis represents the scale, and its  $x$ -axis represents the position parameter  $b$ .

In practice we need to discretize the scale space, and the CWT is computed for scales between  $a_{\min}$  and  $a_{\max}$  with a step  $\delta_a$ .  $a_{\min}$  must be chosen enough large to discretize properly the wavelet function, and  $a_{\max}$  is limited by the number  $N$  of samples in the data. For the experiment shown in Fig. 2,  $a_{\min}$  was set to 0.66 and since the dilated Mexican hat wavelet at scale  $a$  is approximately supported in  $[-4a, 4a]$ , we choose  $a_{\max} = N/8$ . The number of scales  $J$  is defined as the number of voices per octave multiplied by the number of octaves (the number of octaves is the integral part of  $\log_2(a_{\max}/a_{\min})$ ). The number of voices per octave is generally chosen equal to 12, which guaranties a good resolution in scale and



**Numerical Issues When Using Wavelets, Figure 2**

**Top:** 1D signal. **Bottom:** CWT computed with the Mexican Hat wavelet, the y-axis represents the scale and the x-axis represents the position parameter  $b$

the possibility to reconstruct the signal from its wavelet coefficients. We then have  $J = 12 \log_2(a_{\max}/a_{\min})$ , and  $\delta_a = (a_{\max} - a_{\min})/(J - 1)$ .

The CWT algorithm is the following:

- 1: Set the values  $a_{\min}$ ,  $a_{\max}$ ,  $J$ . These values depend on both the chosen wavelet function  $\psi$  and the number of samples  $N$ .
- 2: Set  $\delta_a = (a_{\max} - a_{\min})/(J - 1)$  and  $a = a_{\min}$ .
- 3: for  $a = a_{\min}$  to  $a_{\max}$  with step  $\delta_a$  do
  - Compute  $\psi_a = \psi(x/a)/\sqrt{a}$ .
  - Convolve the input data  $D$  with  $\bar{\psi}_a$  to get  $W(a, \cdot) = 1/\sqrt{a}(\bar{\psi}_a * D)$ . The convolution product can be done either in the direct space or in the Fourier space.
  - $a = a + \delta_a$
- 4:  $W$  contains the CWT of  $D$ .

If the convolution is performed in the Fourier space (i.e.  $\psi_a * D = \text{IFFT}(\text{FFT}(\psi_a)\text{FFT}(D))$ , where FFT and IFFT denote respectively the Fourier transform and its inverse), the data is assumed to be periodic. In this case, the computation of the CWT requires  $O(12N(\log_2 N)^2)$

operations [26]. If the convolution is done in the direct space, we can choose other ways to deal with the borders. For instance, we may prefer to consider mirror reflexive boundary conditions (i.e. for  $k = 0, \dots, N - 1$  we have  $D(-k) = D(k)$  and  $D(N + k) = D(N - 1 - k)$ ).

The choice of the wavelet function is let to the user. As described above, the only constraint is to have a function with a zero mean (admissibility condition). Hence, a large class of functions verifies it and we can adapt the analyzing tool, i.e. the wavelet, to the data. For oscillating data such as audio signals or seismic data, we will prefer a wavelet function which oscillates like the Morlet wavelet. While for other kind of data such as spectra, it is better to choose a wavelet function with minimum oscillation and the Mexican hat would certainly be a good choice. The wavelet function can also be complex, in which case the wavelet transform will be complex. Both the modulus and the phase will carry information about the data.

Here, we have considered only 1D data. For higher dimensional data, we can apply exactly the same approach as above. For 2D data for example, the wavelet function will be defined as a function of five parameters (position  $(b_x, b_y)$ , scale in the two directions  $(a_x, a_y)$  and orientation  $\theta$ ) and the wavelet transform of an image will be of dimension five. But The required memory and the computation time would not be acceptable in most applications. Considering an isotropic wavelet reduces significantly to only three the dimensionality. A even more efficient approach is the (bi-)orthogonal wavelet transform algorithm.

### The (Bi-)Orthogonal Wavelet Transform

Many discrete wavelet transform algorithms have been developed [26,41]. The most widely-known one is certainly the orthogonal transform, proposed by Mallat [25] and its bi-orthogonal version [13]. Here, we will introduce the bi-orthogonal through the two-channel iterated filter bank framework.

Using the bi-orthogonal wavelet transform, a signal  $s$  can be decomposed as follows:

$$s(x) = \sum_k c_J[k] \tilde{\phi}_{J,l}(x) + \sum_{j=1}^J \sum_k \tilde{\psi}_{j,k}(x) w_j[k] \quad (5)$$

with  $\phi_{j,l}$  and  $\psi_{j,l}$  are the scaled dilated and translated version of  $\phi$  and  $\psi$ , which are respectively the scaling function and the wavelet function.  $J$  is the number of reso-

lution levels used in the decomposition,  $w_j$  the wavelet (or detail) coefficients at scale  $j$ , and  $c_j$  is a coarse or smooth version of the original signal  $s$ . Thus, the algorithm outputs  $J + 1$  subband arrays. The indexing is such that, here,  $j = 1$  corresponds to the finest scale (high frequencies). Coefficients  $c_j[k]$  and  $w_j[k]$  are obtained by means of the analysis filters  $h$  and  $g$ :

$$\begin{aligned} c_{j+1}[l] &= \sum_k h[k - 2l] c_j[k] \\ w_{j+1}[l] &= \sum_k g[k - 2l] c_j[k] \end{aligned} \quad (6)$$

where  $h$  and  $g$  are such that:

$$\begin{aligned} \frac{1}{2} \phi\left(\frac{x}{2}\right) &= \sum_k h[k] \phi(x - k) \\ \frac{1}{2} \psi\left(\frac{x}{2}\right) &= \sum_k g[k] \phi(x - k) \end{aligned} \quad (7)$$

and the reconstruction of the signal is performed with:

$$\begin{aligned} c_j[l] &= 2 \sum_k (\tilde{h}[k + 2l] c_{j+1}[k] + \tilde{g}[k + 2l] w_{j+1}[k]) \end{aligned} \quad (8)$$

where the filters  $\tilde{h}$  and  $\tilde{g}$  must verify the conditions of dealiasing and exact reconstruction:

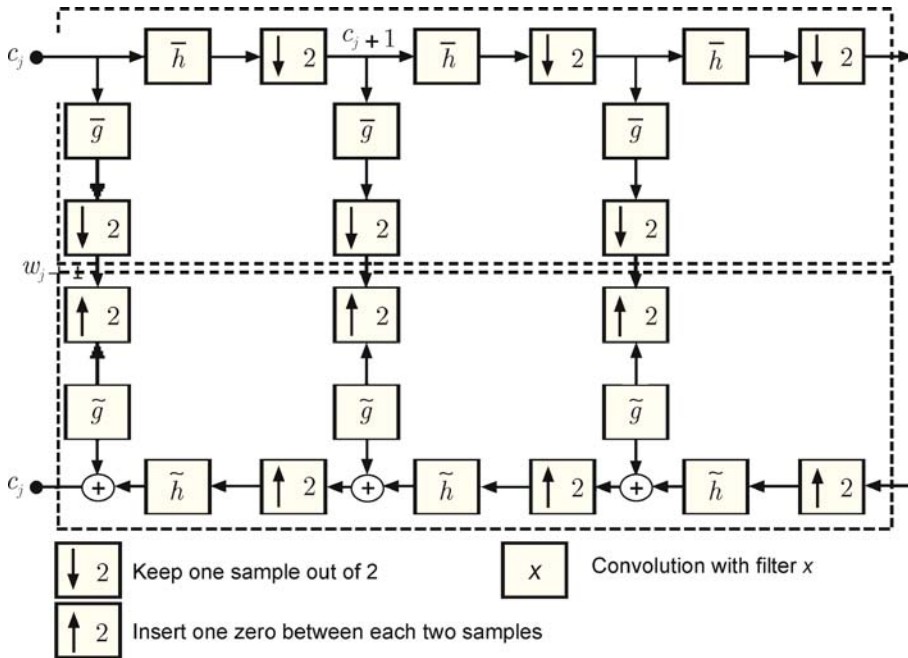
$$\begin{aligned} \hat{h}^*\left(v + \frac{1}{2}\right) \hat{h}(v) + \hat{g}^*\left(v + \frac{1}{2}\right) \hat{g}(v) &= 0 \\ \hat{h}^*(v) \hat{h}(v) + \hat{g}^*(v) \hat{g}(v) &= 1 \end{aligned} \quad (9)$$

or equivalently, in the  $z$ -transform domain:

$$\begin{aligned} H(-z^{-1}) \tilde{H}(z) + G(-z^{-1}) \tilde{G}(z) &= 0 \\ H(z^{-1}) \tilde{H}(z) + G(z^{-1}) \tilde{G}(z) &= 1. \end{aligned}$$

Note that in terms of filter banks, the bi-orthogonal wavelet transform becomes orthogonal when  $h = \tilde{h}$  and  $g = \tilde{g}$ , in which case  $h$  is a conjugate mirror filter.

In the decomposition,  $c_{j+1}$  and  $w_{j+1}$  are computed by successively convolving  $a_j$  with the filters  $\tilde{h}$  (low-pass) and  $g$  (high-pass). Each resulting channel is decimated by suppression of one sample out of two. The high-frequency channel  $w_{j+1}$  is left, and we iterate with the low-frequency part  $c_{j+1}$  (upper part of Fig. 3). In the reconstruction, we restore the sampling by inserting a 0 between each sample, then we convolve with the dual filters  $\tilde{h}$  and  $\tilde{g}$ , we add the resulting coefficients and we multiply the result by 2. The procedure is iterated up to the smallest scale (lower part of Fig. 3).



**Numerical Issues When Using Wavelets, Figure 3**

Fast pyramidal algorithm associated to the bi-orthogonal wavelet transform. *Top*: Fast analysis transform with a cascade of filtering with  $\tilde{h}$  and  $\tilde{g}$  followed by factor 2 subsampling. *Bottom*: Fast inverse transform by progressively inserting zeros and filtering with dual filters  $\tilde{h}$  and  $\tilde{g}$

**Numerical Issues When Using Wavelets, Table 1**  
**7/9 Filter bank (normalized to a unit mass)**

$h$	$g$	$\tilde{h}$	$\tilde{g}$
0	0.02674875741	0.02674875741	0
-0.04563588155	0.0168641184	-0.0168641184	0.04563588155
-0.02877176311	-0.0782232665	-0.0782232665	-0.02877176311
0.295635881557	-0.26686411844	0.26686411844	-0.295635881557
0.557543526229	0.60294901823	0.60294901823	0.557543526229
0.295635881557	-0.26686411844	0.26686411844	-0.295635881557
-0.02877176311	-0.0782232665	-0.0782232665	-0.02877176311
-0.04563588155	0.0168641184	-0.0168641184	0.04563588155
0	0.02674875741	0.02674875741	0

Compared to the CWT, we have much less scales, because we consider only *dyadic scales*, i. e. scales  $a_j$  which are a power of two of the initial scale  $a_0$  ( $a_j = 2^j a_0$ ). Therefore, for a data set with  $N$  samples, we will typically use  $J = \log(N) - 1$  scales. The algorithm is the following:

- 1: Set  $c_0 = D$ ,  $J = \log(N) - 1$ .
- 2: for  $j = 0$  to  $J - 1$  do
  - Compute  $c_{j+1} = \tilde{h} * c_j$ , down-sample by a factor 2.
  - Compute  $w_{j+1} = \tilde{g} * c_j$ , down-sample by a factor 2.
  - $j = j + 1$
- 3: The set  $\mathcal{W} = \{w_1, \dots, w_J, c_J\}$  represents the wavelet transform of the data.

The discrete bi-orthogonal wavelet transform (DWT) is also computationally very efficient, requiring  $O(N)$  operations as compared to  $O(N \log N)$  of the fast Fourier transform ( $N$  is the number of samples in data). The most used filters are certainly the 7/9 filters (by default in the JPEG 2000 norm), which are given in Table 1.

In the literature, the filter bank can be given such that it is normalized to a unit mass  $\sum_k h[k] = 1$ , or to a unit  $\ell_2$ -norm  $\sum_k h[k]^2 = 1$ .

The above DWT algorithm can be easily extended to any dimension by *separable* (tensor) products of a scaling function  $\phi$  and a wavelet  $\psi$ . For instance, the two-dimensional algorithm is based on separate variables leading to prioritizing of horizontal, vertical and diagonal directions. The scaling function is defined by  $\phi(x, y) = \phi(x)\phi(y)$ , and the passage from one resolution to the next is achieved by:

$$\begin{aligned}
 c_{j+1}[k, l] &= \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} h[m - 2k] h[n - 2l] c_j[m, n] \\
 &= (\tilde{h} \tilde{h} * c_j)[k, l]
 \end{aligned}
 \tag{10}$$

The detail signal is obtained from three wavelets:

- Vertical wavelet:  $\psi^1(x, y) = \phi(x) \psi(y)$
- Horizontal wavelet:  $\psi^2(x, y) = \psi(x) \phi(y)$
- Diagonal wavelet:  $\psi^3(x, y) = \psi(x) \psi(y)$

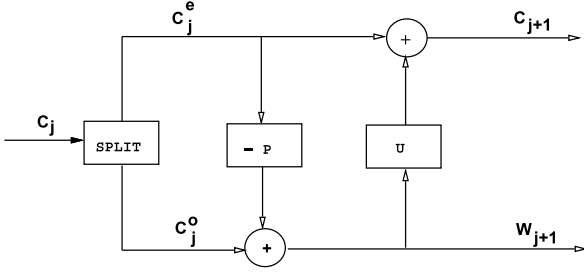
which leads to three wavelet subimages at each resolution level. For three dimensional data, seven wavelet subcubes are created at each resolution level, corresponding to an analysis in seven directions.

For a  $N \times N$  image  $D$ , the algorithm is the following:

- 1: Set  $c_0 = D$ ,  $J = \log(N) - 1$ .
- 2: for  $j = 0$  to  $J - 1$  do
  - Compute  $c_{j+1} = \tilde{h} \tilde{h} * c_j$ , suppress one sample out of two in each dimension.
  - Compute  $w_{j+1}^1 = \tilde{g} \tilde{h} * c_j$ , suppress one sample out of two in each dimension.
  - Compute  $w_{j+1}^2 = \tilde{h} \tilde{g} * c_j$ , suppress one sample out of two in each dimension.
  - Compute  $w_{j+1}^3 = \tilde{g} \tilde{g} * c_j$ , suppress one sample out of two in each dimension.
  - $j = j + 1$
- 3: The set  $\mathcal{W} = \{w_1^1, w_1^2, w_1^3, \dots, w_J^1, w_J^2, w_J^3, c_J\}$  represents the wavelet transform of the data.

### The Lifting Scheme

A lifting is an elementary modification of perfect reconstruction filters, which is used to improve the wavelet properties. The lifting scheme [46] is a flexible technique that has been used in several different settings, for easy construction and implementation of traditional wavelets [46], and for the construction of wavelets on arbitrary domains such as bounded regions of  $\mathbb{R}^d$  (second generation wavelets [45]) or surfaces (spherical wavelets [33]). To optimize the approximation and compression of signals and images, the lifting scheme has also been widely



**Numerical Issues When Using Wavelets, Figure 4**  
The lifting scheme – forward direction

used to construct adaptive wavelet bases with signal-dependent liftings. For example, short wavelets are needed in the neighborhood of singularities, but long wavelets with more vanishing moments improve the approximation of regular regions.

Its principle is to compute the difference between a true coefficient and its prediction:

$$w_{j+1}[l] = c_j[2l+1] - P(c_j[2l-2L], \dots, c_j[2l-2], c_j[2l], c_j[2l+2], \dots, c_j[2l+2L]). \quad (11)$$

A pixel at an odd location  $2l+1$  is then predicted using pixels at even locations.

Computing the wavelet transform using lifting scheme consists of several stages. The idea is to first compute a trivial wavelet transform (the Lazy wavelet) and then improve its properties using alternating lifting and dual lifting steps. The transformation is done in three steps:

1. **Split:** This corresponds to Lazy wavelets which splits the signal into even and odd indexed samples:

$$\begin{aligned} c_j^e[l] &= c_j[2l] \\ c_j^o[l] &= c_j[2l+1]. \end{aligned} \quad (12)$$

2. **Predict:** Calculate the wavelet coefficient  $w_{j+1}[l]$  as the prediction error of  $c_j^o[l]$  from  $c_j^e[l]$  using the prediction operator  $P$ :

$$w_{j+1}[l] = c_j^o[l] - P(c_j^e[l]). \quad (13)$$

3. **Update:** The coarse approximation  $c_{j+1}$  of the signal is obtained by using  $c_j^e[l]$  and  $w_{j+1}[l]$  and the update operator  $U$ :

$$c_{j+1}[l] = c_j^e[l] + U(w_{j+1}[l]). \quad (14)$$

The lifting steps are easily inverted by:

$$\begin{aligned} c_j[2l] &= c_j^e[l] = c_{j+1}[l] - U(w_{j+1}[l]) \\ c_j[2l+1] &= c_j^o[l] = w_{j+1}[l] + P(c_j^e[l]). \end{aligned} \quad (15)$$

Some examples of wavelet transforms via the lifting scheme are:

- **Haar wavelet via lifting:** the Haar transform can be performed via the lifting scheme by taking the predict operator equal to the identity, and an update operator which halves the difference. The transform becomes:

$$\begin{aligned} w_{j+1}[l] &= c_j^o[l] - c_j^e[l] \\ c_{j+1}[l] &= c_j^e[l] + \frac{w_{j+1}[l]}{2}. \end{aligned}$$

All computation can be done in-place.

- **Linear wavelets via lifting:** the identity predictor used before is correct when the signal is constant. In the same way, we can use a linear predictor which is correct when the signal is linear. The predictor and update operators are now:

$$\begin{aligned} P(c_j^e[l]) &= \frac{1}{2} (c_j^e[l] + c_j^e[l+1]) \\ U(w_{j+1}[l]) &= \frac{1}{4} (w_{j+1}[l-1] + w_{j+1}[l]). \end{aligned}$$

It is easy to verify that:

$$\begin{aligned} c_{j+1}[l] &= -\frac{1}{8} c_j[2l-2] + \frac{1}{4} c_j[2l-1] + \frac{3}{4} c_j[2l] \\ &\quad + \frac{1}{4} c_j[2l+1] - \frac{1}{8} c_j[2l+2] \end{aligned}$$

which is the bi-orthogonal Cohen–Daubechies–Feauveau [12] wavelet transform.

The lifting factorization of the popular (7/9) filter pair leads to the following implementation [14]:

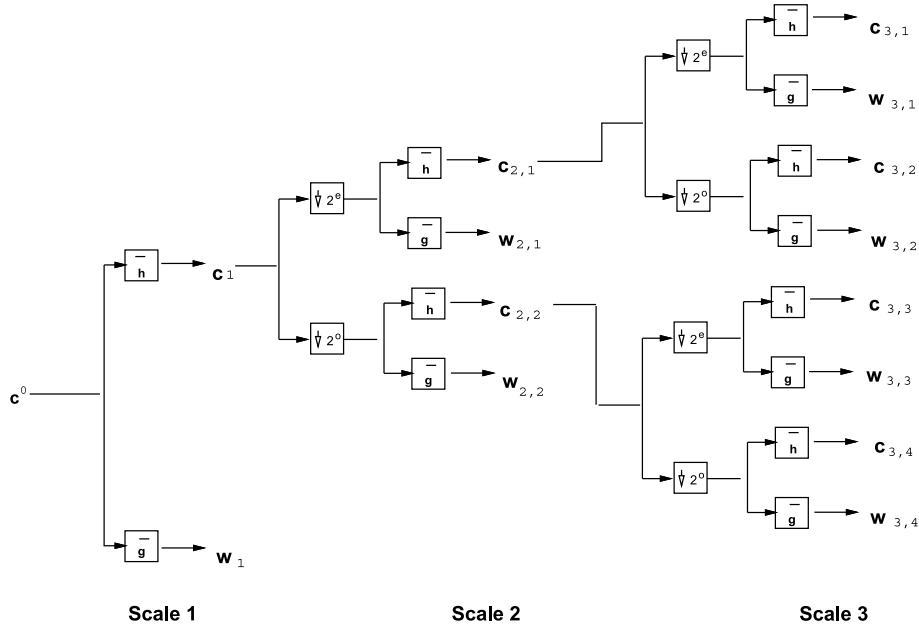
$$\begin{aligned} s^{(0)}[l] &= c_j[2l] \\ d^{(0)}[l] &= c_j[2l+1] \\ d^{(1)}[l] &= d^{(0)}[l] + \alpha (s^{(0)}[l] + s^{(0)}[l+1]) \\ s^{(1)}[l] &= s^{(0)}[l] + \beta (d^{(1)}[l] + d^{(1)}[l-1]) \\ d^{(2)}[l] &= d^{(1)}[l] + \gamma (s^{(1)}[l] + s^{(1)}[l+1]) \\ s^{(2)}[l] &= s^{(1)}[l] + \delta (d^{(2)}[l] + d^{(2)}[l-1]) \\ c_{j+1}[l] &= u s^{(2)}[l] \\ w_{j+1}[l] &= \frac{d^{(2)}[l]}{u} \end{aligned} \quad (16)$$

with

$$\begin{aligned} \alpha &= -1.586134342 \\ \beta &= -0.05298011854 \\ \gamma &= 0.8829110762 \\ \delta &= 0.4435068522 \\ u &= 1.149604398. \end{aligned} \quad (17)$$

Every wavelet transform can be written via lifting.





**Numerical Issues When Using Wavelets, Figure 5**  
1D undecimated wavelet transform

### Integer Wavelet Transform

When the input data consist of integer values, the wavelet transform is not necessarily integer-valued. For lossless coding and compression, it is useful to have a wavelet transform which produces integer values. We can build an integer version of every wavelet transform [7]. For instance, denoting  $\lfloor x \rfloor$  as the largest integer not exceeding  $x$ , the integer Haar transform (also called “S” transform) can be calculated by:

$$\begin{aligned} w_{j+1}[l] &= c_j^o[l] - c_j^e[l] \\ c_{j+1}[l] &= c_j^e[l] + \left\lfloor \frac{w_{j+1}[l]}{2} \right\rfloor \end{aligned} \quad (18)$$

while the reconstruction is

$$\begin{aligned} c_j[2l] &= c_{j+1}[l] - \left\lfloor \frac{w_{j+1}[l]}{2} \right\rfloor \\ c_j[2l+1] &= w_{j+1}[l] + c_j[2l]. \end{aligned} \quad (19)$$

More generally, the lifting operators for an integer version of the wavelet transform are:

$$\begin{aligned} P(c_j^e[l]) &= \left\lfloor \sum_k p[k] c_j^e[l-k] + \frac{1}{2} \right\rfloor \\ U(w_{j+1}[l]) &= \left\lfloor \sum_k u[k] w_{j+1}[l-k] + \frac{1}{2} \right\rfloor \end{aligned} \quad (20)$$

where  $p$  and  $u$  are appropriate filters associated to primal and dual lifting steps.

For instance, the linear integer wavelet transform<sup>2</sup> is given by

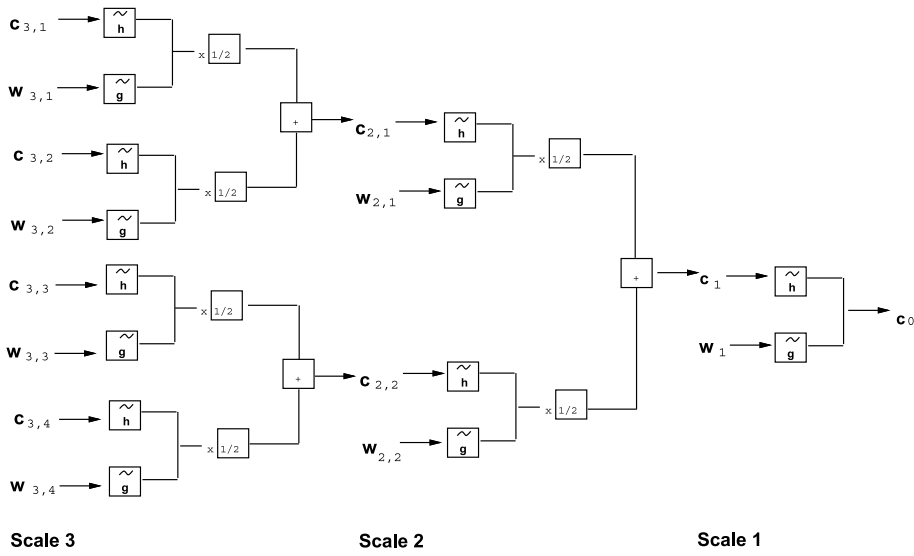
$$\begin{aligned} w_{j+1}[l] &= c_j^o[l] - \left\lfloor \frac{1}{2} (c_j^e[l] + c_j^e[l+1]) + \frac{1}{2} \right\rfloor \\ c_{j+1}[l] &= c_j^e[l] + \left\lfloor \frac{1}{4} (w_{j+1}[l-1] + w_{j+1}[l]) + \frac{1}{2} \right\rfloor. \end{aligned} \quad (21)$$

More filters can be found in [7]. In lossless compression of integer-valued digital images, even if there is no filter that consistently performs better than all the other filters on all images, it was observed that the linear integer wavelet transform performs generally better than other integer wavelet transforms using other filters [7].

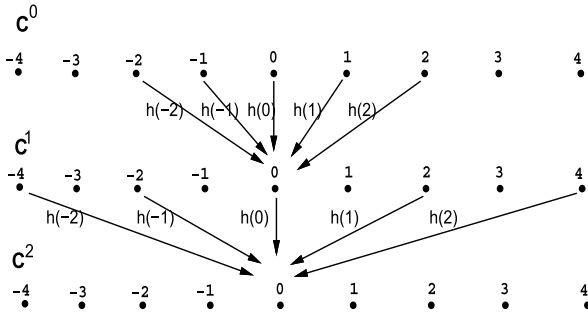
### The Undecimated Wavelet Transform

The undecimated wavelet transform, UWT, consists of keeping the filter bank construction which provides a fast and dyadic algorithms, but eliminating the decimation step in the orthogonal wavelet transform [17,22]:  $c_1 = \tilde{h} * c_0$  and  $w_1 = \tilde{g} * c_0$ . By separating even and odd

<sup>2</sup>This integer wavelet transform is based upon a symmetric, bi-orthogonal wavelet transform built from the interpolating Deslauriers–Dubuc scaling function where both the high-pass filter and its dual has 2 vanishing moments [26].



Numerical Issues When Using Wavelets, Figure 6  
1D undecimated wavelet reconstruction



Numerical Issues When Using Wavelets, Figure 7  
Passage from  $c_0$  to  $c_1$ , and from  $c_1$  to  $c_2$  with the UWT à trous algorithm

pixels in  $c_1$  and  $w_1$ , we get  $(c_1^E, w_1^E)$  and  $(c_1^O, w_1^O)$ , and both parts obviously allow us to reconstruct perfectly  $c_0$ . The reconstruction can be obtained by

$$c_0 = \frac{1}{2} (\tilde{h} * c_1^E + \tilde{g} * w_1^E + \tilde{h} * c_1^O + \tilde{g} * w_1^O). \quad (22)$$

For the passage to the next resolution, both  $c_1^E$  and  $c_1^O$  are decomposed, leading, after the splitting into even and odd pixels, to four coarse arrays associated with  $c_2$ . All of the four data sets can again be decomposed in order to obtain the third decomposition level, and so on.

Figure 5 shows the 1D UWT (UWT) decomposition. The decimation step is not applied and both  $w_1$  and  $c_1$  have the same size as  $c_0$ .  $c_1$  is then split into  $c_1^E$  (even pixels) and  $c_1^O$  (odd pixels), and the same decomposition is applied to both  $c_1^E$  and  $c_1^O$ .  $c_1^E$  produces  $c_{2,1}$  and  $w_{2,1}$ , while  $c_1^O$  produces  $c_{2,2}$  and  $w_{2,2}$ .  $w_2 = \{w_{2,1}, w_{2,2}\}$  contains the wavelet

coefficients at the second scale, and is also of the same size as  $c_0$ . Figure 6 shows the 1D UWT reconstruction.

It is clear that this approach is much more complicated than the decimated bi-orthogonal wavelet transform. There exists, however, a very efficient way to implement it, called the “à trous” algorithm (“à trous”, a French term, meaning *with holes*) [22,34].  $c_{j+1}[L]$  and  $w_{j+1}[L]$  can be expressed as

$$\begin{aligned} c_{j+1}[L] &= (\tilde{h}^{(j)} * c_j)[L] = \sum_k h[k] c_j[L + 2^j k] \\ w_{j+1}[L] &= (\tilde{g}^{(j)} * c_j)[L] = \sum_k g[k] c_j[L + 2^j k], \end{aligned} \quad (23)$$

where  $h^{(j)}[L] = h[L]$  if  $L/2^j$  is an integer and 0 otherwise. For example, we have

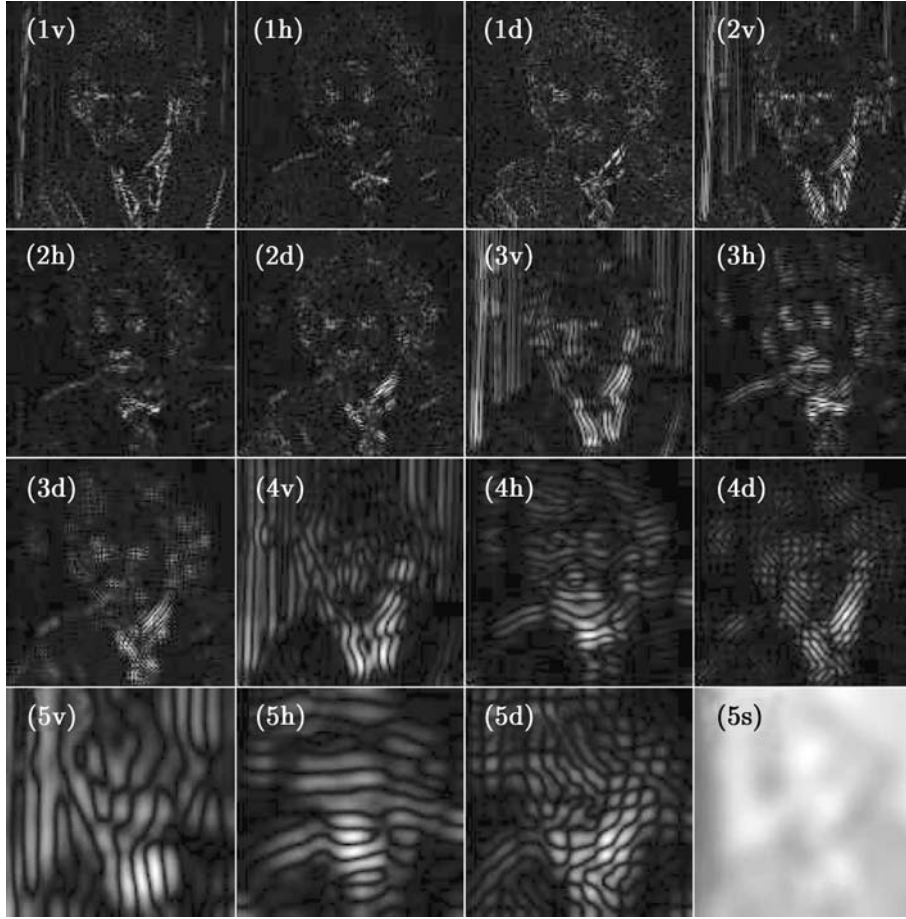
$$h^{(1)} = (\dots, h[-2], 0, h[-1], 0, h[0], 0, h[1], 0, h[2], \dots).$$

The reconstruction is obtained by

$$c_j[L] = \frac{1}{2} [(\tilde{h}^{(j)} * c_{j+1})[L] + (\tilde{g}^{(j)} * w_{j+1})[L]]. \quad (24)$$

The filter bank  $(h, g, \tilde{h}, \tilde{g})$  needs only to verify the exact reconstruction condition written in the  $z$ -transform domain:

$$H(z^{-1}) \tilde{H}(z) + G(z^{-1}) \tilde{G}(z) = 1. \quad (25)$$



**Numerical Issues When Using Wavelets, Figure 8**  
Undecimated wavelet transform of the *Einstein* image

This provides us with a higher degree of freedom when designing the synthesis prototype filter bank.

The à trous algorithm can be extended to 2D, by:

$$\begin{aligned}
 c_{j+1}[k, l] &= (\tilde{h}^{(j)} \tilde{h}^{(j)} * c_j)[k, l] \\
 w_{j+1}^1[k, l] &= (\tilde{g}^{(j)} \tilde{h}^{(j)} * c_j)[k, l] \\
 w_{j+1}^2[k, l] &= (\tilde{h}^{(j)} \tilde{g}^{(j)} * c_j)[k, l] \\
 w_{j+1}^3[k, l] &= (\tilde{g}^{(j)} \tilde{g}^{(j)} * c_j)[k, l].
 \end{aligned} \tag{26}$$

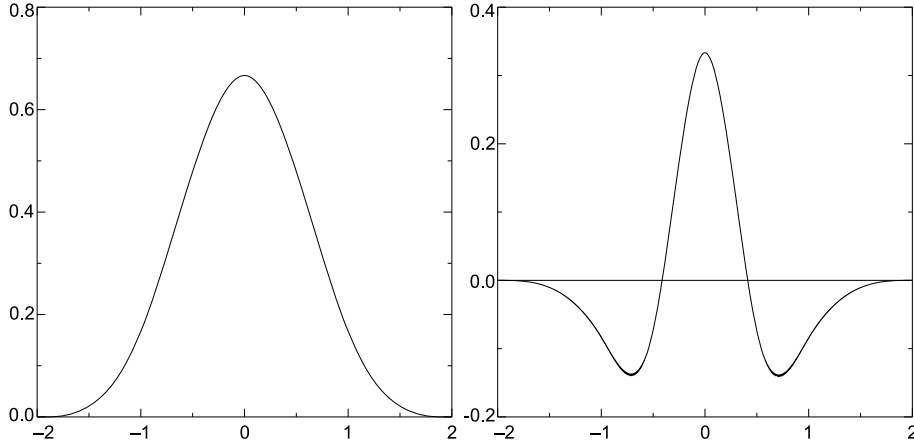
where  $hg * c$  is the convolution of  $c$  by the separable filter  $hg$  (i. e. convolution first along the columns by  $h$  and then convolution along the rows by  $g$ ). At each scale, we have three wavelet images,  $w^1, w^2, w^3$ , and each has the same size as the original image. The redundancy factor is therefore  $3(J - 1) + 1$  [26].

### The 2D Isotropic Undecimated Wavelet Transform

The Isotropic Undecimated Wavelet Transform, IUWT, algorithm is well known in the astronomical domain, because it is well adapted to astronomical data where objects are more or less isotropic in most cases [40]. Requirements for a good analysis of such data are:

- Filters must be symmetric ( $\tilde{h}[k] = h[k]$ , and  $\tilde{g}[k] = g[k]$ ).
- In 2D or higher dimension,  $h, g, \psi, \phi$  must be nearly isotropic.

Filters do not need to be orthogonal or bi-orthogonal and this lack of the need for orthogonality or bi-orthogonality is beneficial for design freedom. For computational reasons, we also prefer to have the separability;  $h[k, l] = h[k]h[l]$ . Separability is not a required condi-



**Numerical Issues When Using Wavelets, Figure 9**

**Left:** the cubic spline function  $\phi$ . **Right:** the wavelet  $\psi$ .  $\psi(x)$  is the difference between two resolutions

tion, but it allows us to have a fast calculation, which is important for a large data set.

This has motivated the following choice for the analysis scaling and wavelet functions [40]:

$$\begin{aligned}\phi_1(x) &= \frac{1}{12}(|x-2|^3 - 4|x-1|^3 + 6|x|^3 \\ &\quad - 4|x+1|^3 + |x+2|^3) \\ \phi(x, y) &= \phi_1(x) \phi_1(y) \\ \frac{1}{4} \psi\left(\frac{x}{2}, \frac{y}{2}\right) &= \phi(x, y) - \frac{1}{4} \phi\left(\frac{x}{2}, \frac{y}{2}\right)\end{aligned}\quad (27)$$

where  $\phi_1(x)$  is the spline of order 3, and the wavelet function is defined as the difference between two resolutions. The related filters  $h$  and  $g$  are defined by:

$$\begin{aligned}h^{(1D)}[k] &= \frac{[1, 4, 6, 4, 1]}{16}, \quad k = -2, \dots, 2 \\ h[k, l] &= h^{(1D)}[k] h^{(1D)}[l] \\ g[k, l] &= \delta[k, l] - h[k, l]\end{aligned}\quad (28)$$

where  $\delta$  is defined as  $\delta[0, 0] = 1$  and  $\delta[k, l] = 0$  for all  $(k, l)$  different from  $(0, 0)$ .

The following useful properties characterize any pair of even-symmetric analysis FIR (finite impulse response) filters  $(h, g = \delta - h)$  such as those of Eq. (28):

**Property 1** For any pair of even symmetric filters  $h$  and  $g$  such that  $g = \delta - h$ , the following holds:

- (i) This FIR filter bank implements a frame decomposition, and perfect reconstruction using FIR filters is possible.
- (ii) The above filters can not implement a tight frame decomposition.

See [39] for a proof.

Figure 9 shows respectively the cubic B-spline scaling function  $\phi$  and the wavelet  $\psi$ .

From the structure of  $g$ , it is easily seen that the wavelet coefficients are obtained just by taking the difference between two resolutions:

$$w_{j+1}[k, l] = c_j[k, l] - c_{j+1}[k, l] \quad (29)$$

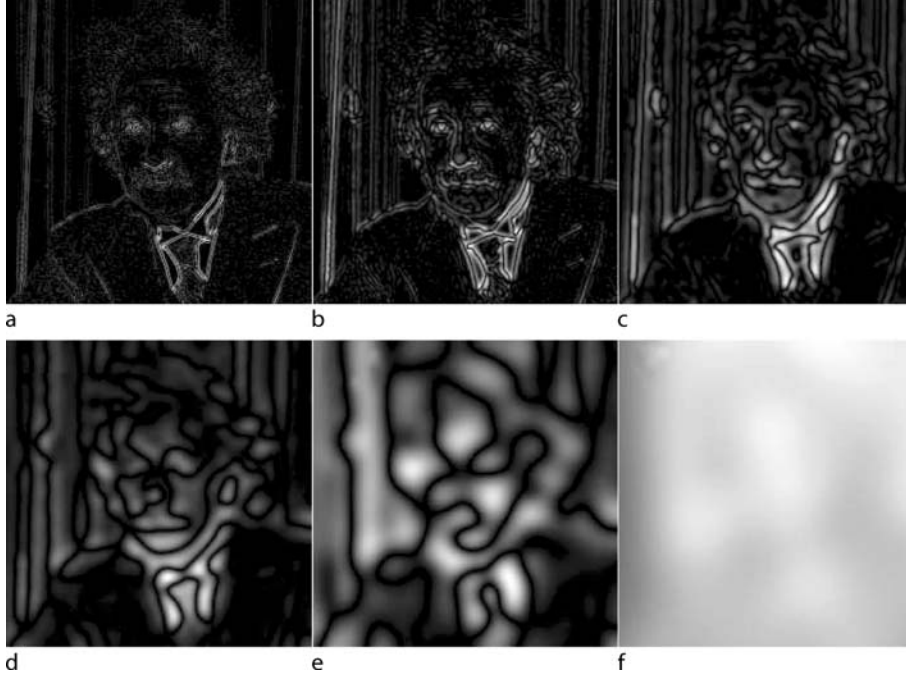
where  $c_{j+1}[k, l] = (\tilde{h}^{(j)} \tilde{h}^{(j)} * c_j)[k, l]$ . At each scale  $j$ , we obtain one subband  $\{w_j\}$  (and not three as in the undecimated WT, denoted UWT above) which has the same number of pixels as the input image.

The reconstruction is obtained by a simple co-addition of all wavelet scales and the final smoothed array, namely

$$c_0[k, l] = c_J[k, l] + \sum_{j=1}^J w_j[k, l]. \quad (30)$$

That is, the synthesis filters are  $\tilde{h} = \delta$  and  $\tilde{g} = \delta$ , which are indeed FIR as expected from Property 1(i). This wavelet transformation is very well adapted to the analysis of images which contain isotropic objects such as in astronomy [40] or in biology [19]. This construction has a close relation to the Laplacian pyramidal construction introduced by Burt and Adelson [6] or the FFT-based pyramidal wavelet transform [41].

Figure 10 shows the undecimated isotropic wavelet transform of the image *Einstein* using six resolution levels. This transformation contains 6 bands, each one being of the same size as the original image. The redundancy factor is therefore equal to 6. The simple addition of these six images reproduce exactly the original image.



Numerical Issues When Using Wavelets, Figure 10

Undecimated isotropic wavelet transform of the *Einstein* image. The addition of these six images reproduce exactly the original image

### Relation Between the UWT and the IUWT

Since the dealiasing filter bank condition is not required anymore in the UWT decomposition, we can build the standard three-directional undecimated filter bank using the non-(bi-)orthogonal “Astro” filter bank ( $h^{1D} = [1, 4, 6, 4, 1]/16$ ,  $g^{1D} = \delta - h^{1D} = [-1, -4, 10, -4, -1]/16$  and  $\tilde{h} = \tilde{g} = \delta$ ). In two dimensions, this filter bank leads to a wavelet decomposition with three orientations  $w_j^1, w_j^2, w_j^3$  at each scale  $j$ , but with the same property as for the IUWT, i. e. the sum of all scales reproduces the original image:

$$c_0[k, l] = c_j[k, l] + \sum_{j=1}^J \sum_{d=1}^3 w_j^d[k, l]. \quad (31)$$

Indeed, a straightforward calculation immediately shows that [39]:

$$w_j^1 + w_j^2 + w_j^3 = c_j - c_{j+1}. \quad (32)$$

Therefore, the sum of the three directions reproduces the IUWT detail band at scale  $j$ . Figure 11 shows the UWT of the galaxy NGC2997. When we add the three directional wavelet bands at a given scale, we recover exactly the isotropic undecimated scale. When we add all bands, we recover exactly the original image. The relation between the two undecimated decompositions is clear.

### Designing Non-orthogonal Filter Banks

#### A Surprising Result

Because the decomposition is non-subsampled, there are many ways to reconstruct the original image from its wavelet transform<sup>3</sup>. For a given filter bank  $(h, g)$ , any filter bank  $(\tilde{h}, \tilde{g})$  which satisfies the reconstruction condition of Eq. (25) leads to exact reconstruction. For instance, for isotropic  $h$ , if we choose  $\tilde{h} = h$ . (the synthesis scaling function  $\tilde{\phi} = \phi$ ) we obtain a filter  $\tilde{g}$  defined by [39]:

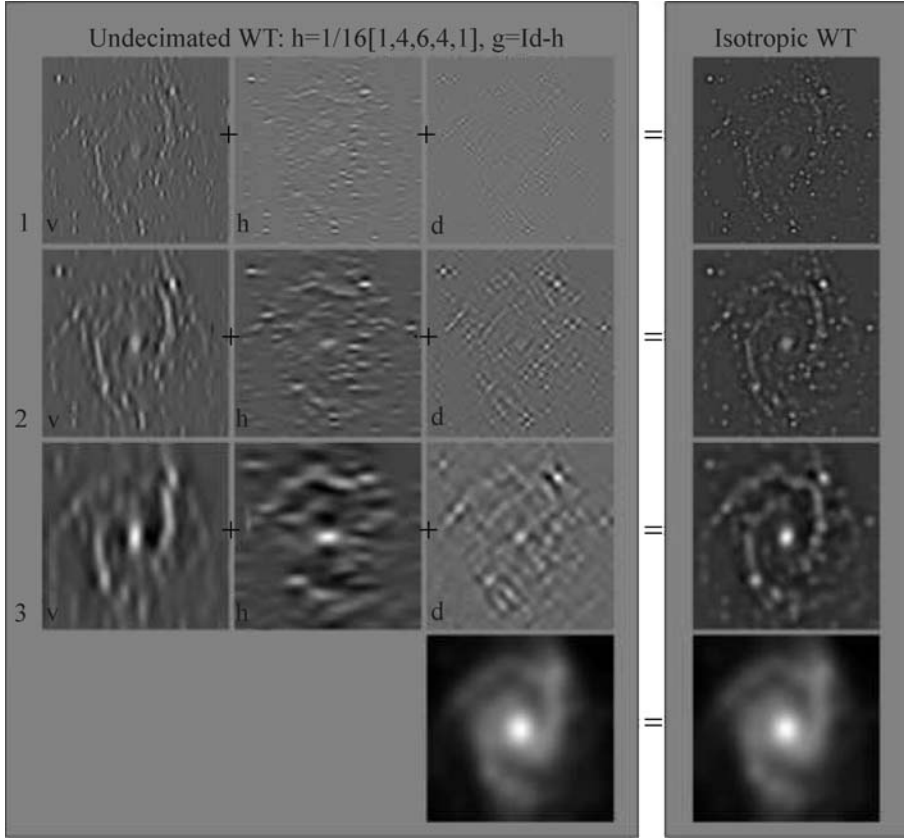
$$\tilde{g} = \delta + h$$

Again, as expected from Property 1, the analysis filter bank  $(h, g = \delta - h)$  implements a (non-tight) frame decomposition for FIR symmetric  $h$ , where  $\tilde{h} = h$  and  $\tilde{g} = \delta + h$  are also FIR filters. For instance, if  $h = [1, 4, 6, 4, 1]/16$ , then  $\tilde{g} = [1, 4, 22, 4, 1]/16$ .  $\tilde{g}$  is positive [39]. This means that  $\tilde{g}$  is no longer related to a wavelet function. The synthesis scaling function related to  $\tilde{g}$  is defined by:

$$\frac{1}{2} \tilde{\psi}\left(\frac{x}{2}\right) = \phi(x) + \frac{1}{2} \phi\left(\frac{x}{2}\right). \quad (33)$$

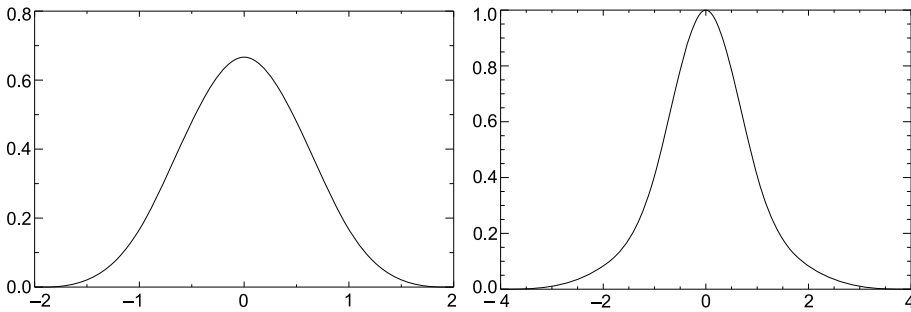
<sup>3</sup>In frame theory parlance, we would say that the UWT frame synthesis operator is not injective.





Numerical Issues When Using Wavelets, Figure 11

UWT of the galaxy NGC2997 using the Astro filter bank. The addition of three bands at a given scale is exactly the band related to the isotropic wavelet transform. Addition of all bands reproduces exactly the original image



Numerical Issues When Using Wavelets, Figure 12

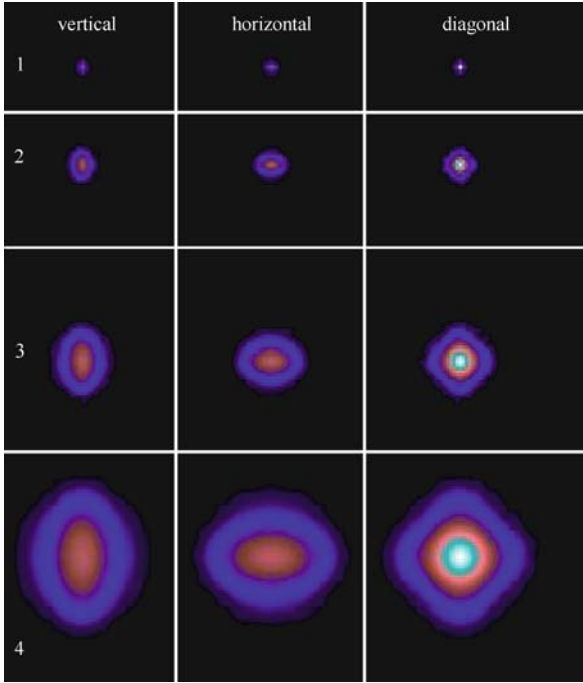
Left: the  $\tilde{\phi}$  synthesis scaling function. Right: the  $\tilde{\psi}$  detail synthesis function

Finally, note that choosing  $\tilde{\phi} = \phi$ , any synthesis function  $\tilde{\psi}$  which satisfies

$$\hat{\psi}(2\nu) \hat{\psi}(2\nu) = \hat{\phi}^2(\nu) - \hat{\phi}^2(2\nu) \quad (34)$$

leads to an exact reconstruction [26] and  $\hat{\psi}(0)$  can take any value. The synthesis function  $\tilde{\psi}$  does not need to verify the admissibility condition (i. e. to have a zero mean).

Figure 12 shows the two scaling functions  $\tilde{\phi}(x) (= \phi)$  and  $\tilde{\psi}(x)$  used in the reconstruction in 1D, corresponding to the synthesis filters  $\tilde{h} = h$  and  $\tilde{g} = \delta + h$ . Figure 13 shows the backprojection of a wavelet coefficient in 2D (all wavelet coefficients are set to zero, except one), when the non-zero coefficient belongs to different bands. We can see that the reconstruction functions are positive.



**Numerical Issues When Using Wavelets, Figure 13**

**Back projection:** Each image corresponds to the backprojection of one wavelet coefficient. All these reconstructed images are positive (no negative values). From left to right, the coefficient belongs to the vertical, horizontal and diagonal direction. From top to bottom, the scale index increases

Finally, we have an expansion of a 1D signal  $s$ ,

$$s(x) = \sum_k c_j[k] \tilde{\phi}_{j,k}(x) + \sum_{j=1}^J \sum_k w_j[k] \tilde{\psi}_{j,k}(x) \quad (35)$$

where  $\tilde{\phi}$  and  $\tilde{\psi}$  are not wavelet functions (both of them have a non-zero mean and are positive), but the  $w_j$  are wavelet coefficients.

### Reconstruction from the Haar Undecimated Coefficients

The Haar filters ( $h = \tilde{h} = [1/2, 1/2]$ ,  $g = \tilde{g} = [-1/2, 1/2]$ ) are not considered as good filters in practice because of their lack of smoothness. They are however very useful in many situations such as denoising where their simplicity allows us to derive analytical or semi-analytical detection levels even when the noise does not follow a Gaussian distribution.

Adopting the same design approach as before, we can reconstruct a signal from its Haar wavelet coefficients choosing a smooth scaling function. For instance,

if  $\tilde{h} = [1, 4, 6, 4, 1]/16$ , it is easy to derive that the  $z$  transforms of these three filters are respectively:

$$\begin{aligned} H(z) &= \frac{1 + z^{-1}}{2}, \\ G(z) &= \frac{z^{-1} - 1}{2}, \\ \tilde{H}(z) &= \frac{z^2 + 4z + 6 + 4z^{-1} + z^{-2}}{16}. \end{aligned} \quad (36)$$

From the exact reconstruction condition in Eq. (25), we obtain:

$$\tilde{G}(z) = \frac{1 - \tilde{H}(z)H(z^{-1})}{G(z^{-1})}. \quad (37)$$

In the case of the spline filter bank, this yields after some re-arrangement (where we used simple convolution properties of splines),

$$\begin{aligned} \tilde{G}(z) &= -2 \frac{1 - z^3 \left( \frac{1+z^{-1}}{2} \right)^5}{1 - z^{-1}} \\ &= z^3 \frac{1 + 6z^{-1} + 16z^{-2} - 6z^{-3} - z^{-4}}{16} \end{aligned} \quad (38)$$

which is the  $z$ -transform of the corresponding filter  $\tilde{g} = [1, 6, 16, -6, -1]/16$ .

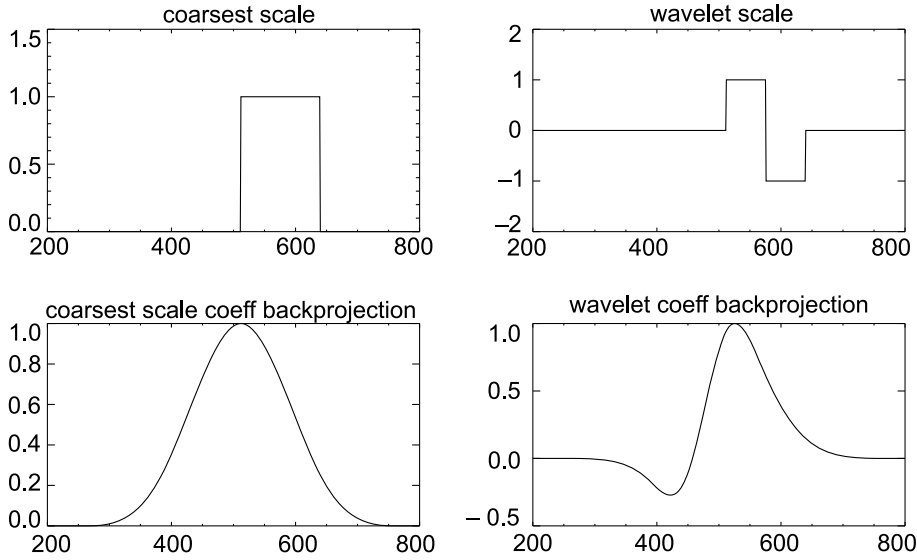
The Haar analysis filters fulfill the following property:

**Property 2** *Haar analysis filters can implement a tight frame expansion (more precisely, one scale of the Haar wavelet UWT does). Perfect reconstruction with FIR synthesis filters is possible.*

Figure 14, upper left and right, depicts the coarsest scale and a wavelet scale of the Haar transform when the input signal contains only zero values except one sample (Dirac  $\delta[k]$ ). Figure 14, bottom left, portrays the backprojection of a Dirac at the coarsest scale (all coefficients are set to zero) and Fig. 14, bottom right, shows the backprojection of a Haar wavelet coefficient. Since the synthesis filters are regular, the backprojection of a Dirac does not produce any block staircase-like artifact. Finally, we would like to point out that other alternatives exist. For example the filter bank ( $h = [1/2, -1/2]$ ,  $g = [-1/4, 1/2, -1/4]$ ,  $\tilde{h} = [1, 3, 3, 1]/8$  and  $\tilde{g} = [1, 6, 1]/4$ ) leads also to an interesting solution where the synthesis filters are both positive.

### Another Interesting Filter Bank

A particular case is obtained when  $\hat{\phi} = \hat{\psi}$  and  $\hat{\psi}(2\nu) = (\hat{\phi}^2(\nu) - \hat{\phi}^2(2\nu))/\hat{\phi}(\nu)$ , which leads to a filter  $\tilde{g}$  equal to  $\delta - h * h$ . In this case, the synthesis function  $\tilde{\psi}$



Numerical Issues When Using Wavelets, Figure 14

Haar Undecimated Transform: *Upper left*: coarsest scale when the signal is  $\delta[k]$ . *Upper right*: one wavelet scale of the Dirac decomposition. *Bottom left*: backprojection of a Dirac at the coarsest scale. *Bottom right*: backprojection of a Haar wavelet coefficient

is defined by  $1/2 \tilde{\psi}(x/2) = \phi(x)$  and the filter  $\tilde{g} = \delta$  is the solution to Eq. (25). We end up with a synthesis scheme where only the smooth part is convolved during the reconstruction. Furthermore, for a symmetric FIR filter  $h$ , it can be easily shown that this filter bank fulfills the statements of Property 1.

Deriving  $h$  from a spline scaling function, for instance  $B_1$  ( $h_1 = [1, 2, 1]/4$ ) or  $B_3$  ( $h_3 = [1, 4, 6, 4, 1]/16$ ) (note that  $h_3 = h_1 * h_1$ ), since  $h$  is even-symmetric (i. e.  $H(z) = H(z^{-1})$ ), the  $z$ -transform of  $g$  is then:

$$\begin{aligned} G(z) &= 1 - H^2(z) \\ &= 1 - z^4 \left( \frac{1 + z^{-1}}{2} \right)^8 \\ &= \frac{-z^4 - 8z^3 - 28z^2 - 56z + 186 - 56z^{-1}}{256} \\ &\quad - \frac{28z^{-2} - 8z^{-3} - z^{-4}}{256} \end{aligned} \quad (39)$$

which is the  $z$ -transform of the filter  $g = [-1, -8, -28, -56, 186, -56, -28, -8, -1]/256$ . We get the following filter bank:

$$\begin{aligned} h &= h_3 = \tilde{h} = \frac{[1, 4, 6, 4, 1]}{16} \\ g &= \delta - h * h \\ &= \frac{[-1, -8, -28, -56, 186, -56, -28, -8, -1]}{256} \end{aligned} \quad (40)$$

$$\tilde{g} = \delta \quad (41)$$

With this filter bank, there is no convolution with the filter  $\tilde{g}$  during the reconstruction. Only the low-pass synthesis filter  $\tilde{h}$  is used. The reconstruction formula is:

$$c_j[l] = (h^{(j)} * c_{j+1})[l] + w_{j+1}[l] \quad (42)$$

and denoting  $L^j = h^{(0)} * \dots * h^{(j-1)}$  and  $L^0 = \delta$ , we have

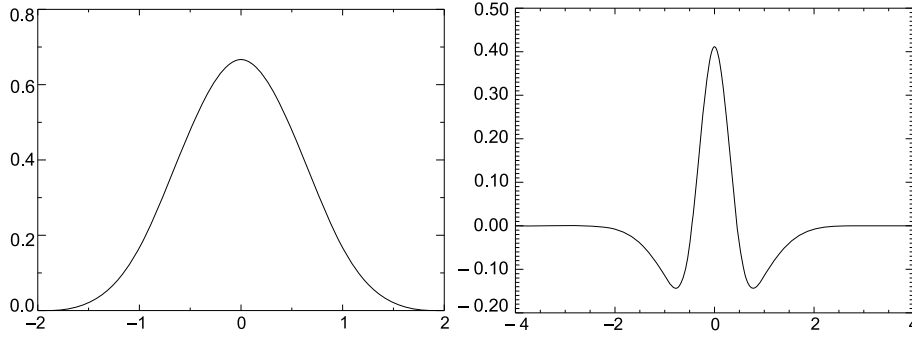
$$c_0[l] = (L^J * c_J)[l] + \sum_{j=1}^J (L^{j-1} * w_j)[l]. \quad (43)$$

Each wavelet scale is convolved with a low-pass filter.

Figure 15 shows the analysis scaling and wavelet functions. The synthesis functions  $\tilde{\phi}$  and  $\tilde{\psi}$  are the same as those in Fig. 12.

### Iterative Reconstruction

Denoting  $\mathcal{W}$  the undecimated wavelet transform operator and  $\mathcal{R}$  the reconstruction operator, and thanks to the exact reconstruction formulae, we have the relation:  $\alpha_S = \mathcal{W}\mathcal{R}\alpha_S$ , where  $S$  is a signal or image and  $\alpha_S$  its wavelet coefficients (i. e.  $\alpha_S = \mathcal{W}S$ ). But we loose one fundamental property of the (bi-)orthogonal WT. Indeed, the relation  $\alpha = \mathcal{W}\mathcal{R}\alpha$  is not true for all  $\alpha$  sets. For example, if we set all wavelet coefficients to zero except one at a coarse scale, there is no image such that its UWT would produce a Dirac at a coarse scale. Another way



**Numerical Issues When Using Wavelets, Figure 15**

*Left: the  $\phi$  analysis scaling function. Right: the  $\psi$  analysis wavelet function. The synthesis functions  $\tilde{\phi}$  and  $\tilde{\psi}$  are the same as those in Fig. 12*

to understand this point is to consider the Fourier domain of a given undecimated scale. Indeed, wavelet coefficients  $\alpha_j$  at scale  $j$  obtained using the wavelet transform operator will contain information only localized at a given frequency band. But any modification of the coefficients at this scale, such as a thresholding ( $\alpha_T = \Delta_T(\alpha)$ , where  $\Delta_T$  is the thresholding operator with threshold  $T$  and  $\alpha_T$  are the thresholded coefficients), will introduce some frequency components which should not exist at this scale  $j$ , and we have  $\alpha_T \neq \mathcal{WR}\alpha_T$ .

### Reconstruction from a Subset of Coefficients

Without loss of generality, we consider hereafter the case of 1D signals. If only a subset of coefficients (for instance after thresholding) is different from zero, we would like to reconstruct an image  $\tilde{S}$  such that its wavelet transform reproduces the non-zero wavelet coefficients. This can be cast as an inverse problem. We want to solve the following optimization problem  $\min_{\tilde{S}} \|M(\alpha_T - \mathcal{W}\tilde{S})\|_2^2$  where  $M_j[k]$  is the multiresolution support of  $\alpha$ , i. e.  $M_j[k] = 1$  if the wavelet coefficient  $\alpha_j[k]$  at scale  $j$  and at position  $k$  is different from zero, and  $M_j[k] = 0$  otherwise. A solution can be obtained using the Landweber iterative scheme [35,41]:

$$\tilde{S}^{n+1} = \tilde{S}^n + \mathcal{RM}[\alpha_T - \mathcal{W}\tilde{S}^n]. \quad (44)$$

If the solution is known to be positive, the positivity constraint can be introduced using the following equation:

$$\tilde{S}^{n+1} = P_+(\tilde{S}^n + \mathcal{RM}[\alpha_T - \mathcal{W}\tilde{S}^n]) \quad (45)$$

where  $P_+$  is the projection on the cone of non-negative images. This iterative scheme can also be interpreted in terms of alternating projections onto convex sets (POCS). It has also proven very effective at many tasks such as image approximation and restoration when using the UWT [39].

### Future Directions

For 2D or 3D data set, wavelet bases present some intrinsic limitations, because they are not adapted to the detection of highly anisotropic elements, such as lines or curvilinear structures in an image, or sheets in a cube. Recently, other multiscale systems like curvelets [8,9,15,36] and ridgelets [10] which are very different from wavelet-like systems have been developed. Curvelets and ridgelets take the form of basis elements which exhibit very high directional sensitivity and are highly anisotropic. A digital implementation of both the ridgelet and the curvelet transform for image denoising has been described in [36]. These new data representations, combined with wavelets, have been used in many applications such as denoising [21,32,36], deconvolution [43], contrast enhancement [42], texture analysis [2,38], detection [23], watermarking [49], component separation [37], inpainting [18] or blind source separation [4,5].

To reach higher sparsity levels, the transforms just mentioned with a fixed geometry can be replaced by adaptive representations using an optimized basis. Geometric transforms such as wedgelets [16] or bandlets [24,28] allow to define an adapted multiscale geometry. These transforms perform a non-linear search for an optimal representation. They offer geometrical adaptivity together with fast and stable algorithms. Recently, Mallat [27] proposed a more biologically inspired procedure named the grouplet transform, which defines a multiscale association field by grouping together pairs of wavelet coefficients.

Following Olshausen and Field [31], one can push one step forward the idea of adaptive sparse representation and requires that the dictionary is not fixed but rather optimized to sparsify a set of exemplar signals/images. Such a learning problem corresponds to finding a sparse matrix factorization as exposed in the K-SVD framework [1]. Ex-

PLICIT structural constraints such as translation invariance can also be enforced on the learned dictionary [3,30].

## Bibliography

### Primary Literature

- Aharon M, Elad M, Bruckstein AM (2006) The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans Signal Process* 54(11):4311–4322
- Arivazhagan S, Ganesan L, Kumar TS (2006) Texture classification using curvelet statistical and co-occurrence features. In: *Proceeding of the 18th International Conference on Pattern Recognition (ICPR 2006)*, vol 2. IEEE Computer Society, Los Alamitos, pp 938–941
- Blumensath T, Davies M (2006) Sparse and shift-invariant representations of music. *IEEE Trans Speech Audio Process* 14(1):50–57
- Bobin J, Moudden Y, Starck J-L, Elad M (2006) Morphological diversity and source separation. *IEEE Trans Signal Process* 13(7):409–412
- Bobin J, Starck J-L, Fadili J, Moudden Y (2007) Sparsity, morphological diversity and blind source separation. *IEEE Trans Image Process* 16(11):2662–2674
- Burt P, Adelson A (1983) The Laplacian pyramid as a compact image code. *IEEE Trans Commun* 31:532–540
- Calderbank R, Daubechies I, Sweldens W, Yeo B-L (1998) Wavelet transforms that map integers to integers. *Appl Comput Harmon Anal* 5:332–369
- Candès EJ, Demanet L, Donoho DL, Ying L (2006) Fast discrete curvelet transforms. *SIAM Multiscale Model Simul* 5(3):861–899
- Candès EJ, Donoho DL (1999) Curvelets – a surprisingly effective nonadaptive representation for objects with edges. In: Cohen A, Rabut C, Schumaker LL (eds) *Curve and surface fitting: Saint-Malo 1999*. Vanderbilt University Press, Nashville
- Candès EJ, Donoho DL (1999) Ridgelets: the key to high dimensional intermittency? *Philos Trans R Soc Lond A* 357:2495–2509
- Cohen A (2003) *Numerical analysis of wavelet methods*. Elsevier, Amsterdam
- Cohen A, Daubechies I, Feauveau J (1992) Biorthogonal bases of compactly supported wavelets. *Commun Pure Appl Math* 45:485–560
- Daubechies I (1992) *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics Press, Philadelphia
- Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. *J Fourier Anal Appl* 4:245–267
- Do MN, Vetterli M (2003) Contourlets. In: Stoeckler J, Welland GV (eds) *Beyond wavelets*. Academic Press, New York
- Donoho DL (1999) Wedgelets: nearly-minimax estimation of edges. *Ann Statist* 27:859–897
- Dutilleul P (1989) An implementation of the “algorithme à trous” to compute the wavelet transform. In: Combes JM, Grossmann A, Tchamitchian P (eds) *Wavelets: Time-frequency methods and phase-space*. Springer, New York
- Elad M, Starck J-L, Donoho DL, Querre P (2006) Simultaneous cartoon and texture image inpainting using Morphological Component Analysis (MCA). *J Appl Comput Harmonic Anal* 19:340–358
- Genovesio A, Olivo-Marin J-C (2003) Tracking fluorescent spots in biological video microscopy. In: Conchello J-A, Cogswell CJ, Wilson T (eds) *Three-dimensional and multidimensional microscopy: Image acquisition and processing X*, vol 4964. SPIE Publications, pp 98–105
- Grossmann A, Kronland-Martinet R, Morlet J (1989) Reading and understanding the continuous wavelet transform. In: Combes JM, Grossmann A, Tchamitchian P (eds) *Wavelets: Time-frequency methods and phase-space*. Springer, Berlin, pp 2–20
- Hennenfent G, Herrmann FJ (2006) Seismic denoising with nonuniformly sampled curvelets. *IEEE Comput Sci Eng* 8(3):16–25
- Holschneider M, Kronland-Martinet R, Morlet J, Tchamitchian P (1989) A real-time algorithm for signal analysis with the help of the wavelet transform. In: Combes JM, Grossmann A, Tchamitchian P (eds) *Wavelets: Time-frequency methods and phase-space*. Springer, New York, pp 286–297
- Jin J, Starck J-L, Donoho DL, Aghanim N, Forni O (2005) Cosmological non-Gaussian signatures detection: Comparison of statistical tests. *Eurasip J* 15:2470–2485
- Le Pennec E, Mallat S (2005) Bandelet image approximation and compression. *SIAM Multiscale Modeling Simul* 4(3):992–1039
- Mallat S (1989) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11:674–693
- Mallat S (1998) *A wavelet tour of signal processing*. Academic Press, New York
- Mallat S (2006) Geometrical grouplets. *Appl Comput Harmonic Anal*, submitted
- Mallat S, Peyrè G (2006) Orthogonal bandlet bases for geometric images approximation. *Com Pure Appl Math*, to appear
- MR/1 (2001) Multiresolution image and data analysis software package, Version 3.0. Multi Resolutions Ltd, <http://www.multiresolution.com>
- Olshausen BA (2002) Sparse coding of time-varying natural images. *J Vision* 2(7):130
- Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609
- Saevansson B, Sveinsson J, Benediktsson J (2006) Speckle reduction of SAR images using adaptive curvelet domain. In: *Proceeding of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS 2003)*, vol 6. IEEE Computer Society, Los Alamitos, pp 4083–4085
- Schröder P, Sweldens W (1995) Spherical wavelets: Efficiently representing functions on the sphere. In: *Computer Graphics (SIGGRAPH 95)*. ACM SIGGRAPH Publications, pp 161–172
- Shensa M (1992) Discrete wavelet transforms: Wedding the à trous and Mallat algorithms. *IEEE Trans Signal Process* 40:2464–2482
- Starck J-L, Bijaoui A, Murtagh F (1995) Multiresolution support applied to image filtering and deconvolution. *CVGIP: Graph Model Image Process* 57:420–431
- Starck J-L, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. *IEEE Trans Image Process* 11(6):131–141
- Starck J-L, Elad M, Donoho DL (2004) Redundant multiscale transforms and their application for morphological component analysis. *Adv Imaging Electron Phys* 132:288–345



38. Starck J-L, Elad M, Donoho DL (2005) Image decomposition via the combination of sparse representation and a variational approach. *IEEE Trans Image Process* 14(10):1570–1582
39. Starck J-L, Fadili J, Murtagh F (2007) The undecimated wavelet decomposition and its reconstruction. *IEEE Trans Image Process* 16:297–309
40. Starck J-L, Murtagh F (2002) *Astronomical image and data analysis*. Springer, Berlin
41. Starck J-L, Murtagh F, Bijaoui A (1998) *Image processing and data analysis: The multiscale approach*. Cambridge University Press, Cambridge
42. Starck J-L, Murtagh F, Candès EJ, Donoho DL (2003) Gray and color image contrast enhancement by the curvelet transform. *IEEE Trans Image Process* 12(6):706–717
43. Starck J-L, Nguyen M, Murtagh F (2003) Wavelets and curvelets for image deconvolution: A combined approach. *Signal Process* 83(10):2279–2283
44. Strang G, Nguyen T (1996) *Wavelet and filter banks*. Wellesley-Cambridge Press, Wellesley
45. Sweldens W (1997) The lifting scheme: A construction of second generation wavelets. *SIAM J Math Anal* 29:511–546
46. Sweldens W, Schröder P (1996) Building your own wavelets at home. In: *Wavelets in Computer Graphics*, ACM SIGGRAPH Course notes. ACM SIGGRAPH Publications, pp 15–87
47. Wavelab 802 (2001) Stanford University, [http://www-stat.stanford.edu/~wavelab/index\\_wavelab802.html](http://www-stat.stanford.edu/~wavelab/index_wavelab802.html)
48. Wavelab 805 (2005) Stanford University, <http://www-stat.stanford.edu/~wavelab/>
49. Zhang Z, Huang W, Zhang J, Yu H, Lu Y (2006) Digital image watermark algorithm in the curvelet domain. In: *Proceeding of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2006)*. IEEE Computer Society, Los Alamitos, pp 105–108

## Books and Reviews

- Antoine J-P, Murenzi R, Vanderghyest P, Ali ST (2004) *Two-dimensional wavelets and their relatives*. Cambridge University Press, Cambridge
- Chan T, Shen J (2005) *Image processing and analysis: Variational, Pde, wavelet, and stochastic methods*. Society for Industrial and Applied Mathematics Press, Philadelphia
- Cohen A (2003) *Numerical analysis of wavelet methods*. Elsevier, Amsterdam
- Daubechies I (1992) Ten lectures on wavelets. In: *Proceedings of CBMS-NSF Regional Conference Series in Applied Mathematics Philadelphia 1992*, vol 61. Society for Industrial and Applied Mathematics Press, Philadelphia
- Hubbard BB (1995) *The world according to wavelets: The story of a mathematical technique in the making*. AK Peters, Wellesley
- Jaffard S, Ryan RD, Meyer Y (2005) *Wavelets: Tools for science and technology*. Society for Industrial and Applied Mathematics Press, Philadelphia
- Mallat S (1999) *A wavelet tour of signal processing*. Academic Press, New York
- Meyer Y, Ryan R (1993) *Wavelets: Algorithms and applications*. Society for Industrial and Applied Mathematics Press, Philadelphia
- Starck J-L, Murtagh F (2006) *Astronomical data analysis*, 2nd edn. Springer, Berlin
- Starck J-L, Murtagh F, Bijaoui A (1998) *Image processing and data analysis: The multiscale approach*. Cambridge University Press, Cambridge
- Vetterli M, Kovacevic J (1995) *Wavelets and subband coding*. Prentice-Hall, New Jersey
- Vidakovic B (1999) *Statistical modeling by wavelets*. In: *Wiley Series in probability and statistics*. Wiley, New York
- Walker JS (1999) *A primer on wavelets and their scientific applications*. Chapman and Hall/CRC, Boca Raton